

# A POLYNOMIAL BOUND ON SOLUTIONS OF QUADRATIC EQUATIONS IN FREE GROUPS

IGOR LYSENOK AND ALEXEI MYASNIKOV

ABSTRACT. We provide polynomial upper bounds on the size of a shortest solution for quadratic equations in a free group. A similar bound is given for parametric solutions in the description of solutions sets of quadratic equations in a free group.

## 1. INTRODUCTION

Let  $G$  be a group and  $X$  a countable set of *variables*. An *equation in  $G$*  is an element  $E$  of the free product  $G * F_X$  where  $F_X$  is the free group freely generated by  $X$ . Usually we write equations in the classical form  $E = 1$  where  $E$  is a word over  $G \cup X^{\pm 1}$  representing a reduced product in  $G * F_X$ . Elements of  $G$  occurring in  $E$  are called *coefficients* of an equation  $E = 1$ . A *solution* of  $E = 1$  in  $G$  is a homomorphism  $\alpha : G * F_X \rightarrow G$  such that  $g^\alpha = g$  for every  $g \in G$  (so-called  *$G$ -homomorphism*) and  $E^\alpha = 1$ . Sometimes it is convenient to consider the quotient  $G_E = G * F_X / (E = 1)$  of  $G * F_X$  modulo the normal subgroup generated by  $E$ , called the *equation group of  $E = 1$* . In this case solutions of  $E = 1$  in  $G$  are precisely  $G$ -homomorphisms  $G * F_X \rightarrow G$ . (Observe that if  $E = 1$  has a solution then  $G$  embeds into  $G_E$ .) In general, there are two natural problems concerning equations from a given class  $\mathcal{E}$ . The first one is the famous Diophantine problem: does there exist an algorithm to check whether or not a given equation from  $\mathcal{E}$  has a solution in a group  $G$ . The second problem is to get an effective description of the set of solutions of an equation in  $\mathcal{E}$ .

A word  $E$  and an equation  $E = 1$  are termed *quadratic* if every variable  $x \in X$  occurring in  $E$  (as  $x$  or  $x^{-1}$ ) occurs precisely twice. Quadratic equations form a very special class among equations of a general form in groups. However, they play an important role in several areas of mathematics. This is not very surprising since quadratic equations are naturally related to the topology of compact surfaces (see for example [7, 24] and Section 2 below). Quadratic equations groups naturally appear in JSJ-decompositions of groups (via QH subgroups) and as group actions of dynamical systems (via interval exchange). Moreover, quadratic equations play a fundamental part in general theory of equations in groups and algebraic (Diophantine) geometry over groups [2, 15, 16, 14, 28]. It has been shown by Kharlampovich and Myasnikov [15, 16] that every system of equations in a free group is rationally equivalent to finitely many systems in non-degenerate triangular quasi-quadratic form (NTQ), which gives a precise analog of the elimination theory from the classical algebraic geometry, with proper extension theorems and nice algorithmic properties (for details see surveys [17, 18]). NTQ systems are crucial in Implicit Function Theorems in free groups [14] and Tarski problems [19, 28].

The theory of quadratic equations in a free group  $F$  goes back to the works of Lyndon [20] and Malcev [22], who described, correspondingly, all solutions of the equation  $x^2 y^2 z^2 = 1$  and

---

The first author has been supported by the Russian Foundation for Fundamental Research .

$[x, y] = [a, b]$  (here  $x, y, z \in X$  are variables and  $a, b \in F$  are the generators of  $F$ ). Malcev anticipated several modern techniques, introducing automorphic equivalence of solutions and focusing on minimal solutions in each automorphic orbit. In [30] Wicks gave a decision algorithm for the Diophantine problem for equations of the type  $W = g$  ( $g \in F$ ). He showed that the problem of solving such an equation can be effectively reduced to solving finitely many particular equations in a free monoid with involution (via “Wicks forms”).

In [5], Comerford and Edmunds proved that the Diophantine problem for arbitrary quadratic equations in a free group  $F$  is decidable. Later Comerford and Edmunds [6] and Grigorchuk and Kurchanov [11] completely described solution sets of quadratic equations in free groups (the result will be formulated below in this section).

In 1982, Makanin [23] proved that the Diophantine problem in free groups is solvable, and a few years later Razborov [25] gave a description of solutions of systems of equations in free groups. These two very influential papers shaped the modern theory of equations in groups. Note that the description of solutions sets of quadratic equations given in [6, 11] is a very special case of Razborov’s description for general equations.

Techniques for solving equations in free and related groups were instrumental in solution of some other decision problems in group theory: for example, the isomorphism problem in hyperbolic groups [27, 9], limit groups [3], and toral relatively hyperbolic groups [8].

In view of applications, the principal question concerning equations in groups is the time complexity of decision algorithms. It has been shown by Bormotov, Gilman, and Myasnikov [1] that one-variable equations in a free group admit polynomial time decision algorithms. Ol’shaskniĭ [24] and Grigorchuk and Kurchanov [11] proved that if the number of variables is fixed then the Diophantine problem for quadratic equations in free groups has a decision algorithm polynomial in the sum of the lengths of the coefficients. However, this is as far as one can go in polynomial time. First, Diekert and Robson showed in [10] that the Diophantine problem for quadratic equations in free monoids (semigroups) is NP-hard. Then Kharlampovich, Lysenok, Myasnikov and Touikan proved that the Diophantine problem for quadratic equations in free groups is precisely NP-complete [13]. Nevertheless, a few important questions remained to be open, above all, the question whether decidable quadratic equations in free groups have solutions polynomially bounded in the size of the equation.

The affirmative answer to this question is given by the following theorem. Here a quadratic word  $Q$  is *orientable* if every variable  $x$  occurring in  $Q$  occurs in  $Q$  with two opposite exponents as  $x$  and  $x^{-1}$  and *non-orientable* otherwise, that is, if a variable occurs in  $Q$  twice with the same exponent  $+1$  or  $-1$ .

**Theorem 1.1.** *Let  $Q$  be a quadratic word. If the equation  $Q = 1$  is solvable in a free group  $F_A$  then there exists a solution  $\alpha$  such that for any variable  $x$ ,*

$$|x^\alpha| \leq \begin{cases} Nn(Q)c(Q) & \text{if } Q \text{ is orientable,} \\ Nn(Q)^2c(Q) & \text{if } Q \text{ is non-orientable,} \end{cases}$$

for some constant  $N$ . Here  $n(Q)$  denotes the total number of variables in  $Q$  and  $c(Q)$  the total length of coefficients occurring in  $Q$ . One can take  $N = 40$  for orientable  $Q$  and  $N = 150$  for non-orientable  $Q$ .

If  $Q$  is standard orientable or semi-standard non-orientable (see Definitions 2.2 and 3.4) then there exists  $\alpha$  with a better bound

$$|x^\alpha| \leq \begin{cases} Nc(Q) & \text{if } Q \text{ is orientable,} \\ Nn(Q)c(Q) & \text{if } Q \text{ is non-orientable,} \end{cases}$$

with  $N = 8$  and  $N = 36$  respectively.

In a similar manner, we give a bound on the size of parametric solutions of quadratic equations which participate in the description of their solution sets. To state the result we need to define a concept of a parametric solution. Let  $T$  be an alphabet of *parameters* which is assumed to be disjoint from the alphabet of constants  $A$  and alphabet of variables  $X$ . A *parametric solution* of an equation  $E = 1$  in  $F_A$  is an  $F_A$ -homomorphism  $\eta : F_{A \cup \text{Var}(E)} \rightarrow F_{A \cup T}$  such that  $E^\eta = 1$  where  $\text{Var}(E)$  denotes the set of variables occurring in  $E$ . If  $\eta$  is a parametric solution of  $E = 1$  then for any  $F_A$ -homomorphism  $\psi : F_{A \cup T} \rightarrow F_A$  we get a solution  $\eta\psi$  of  $E = 1$  in the usual sense, a *specialization* of  $\eta$ . Note that we use here a notion of a solution of an equation  $E = 1$  in a free group  $F_A$  which is slightly different from one introduced above. Instead of taking  $F_A$ -homomorphisms of  $F_{A \cup X}$  we restrict them to the free group  $F_{A \cup \text{Var}(E)}$  involving only variables which occur in  $E$ . This provides a more convenient way for describing *all* possible solutions of a given equation  $E = 1$ .

Let  $\text{Stab}(E)$  denote the group of all  $F_A$ -automorphisms  $\phi$  of  $F_{A \cup \text{Var}(E)}$  such that  $E^\phi$  is conjugate to  $E$ . This group acts on the solution set of  $E = 1$  by left multiplications. Hence any parametric solution  $\eta$  produces a whole bunch of solutions in the usual sense, the union of orbits of specializations of  $\eta$ :

$$\text{Sol}(\eta) = \{\phi\eta\omega \mid \phi \in \text{Stab}(E), \omega \in \text{Hom}_{F_A}(F_{A \cup T}, F_A)\}.$$

The above mentioned result of Comerford–Edmunds [6] and Grigorchuk–Kurchanov [11] asserts that for any quadratic equation  $E = 1$  in  $F_A$  there is (and can be effectively produced) a finite set  $\{\eta_i\}$  of *basic* parametric solutions such that the set of all solution of  $E = 1$  is the union  $\cup_i \text{Sol}(\eta_i)$ .

Let  $\eta$  and  $\theta$  be two parametric solutions of the same equation  $E = 1$  in  $F_A$ . Let us say that  $\eta$  is a *generalization* of  $\theta$  if there are an automorphism  $\phi \in \text{Stab}(E)$  and an endomorphism  $\omega \in \text{End}_{F_A}(F_{A \cup T})$  such that  $\theta = \phi\eta\omega$ . Clearly, in this case we have  $\text{Sol}(\eta) \supseteq \text{Sol}(\theta)$ .

**Theorem 1.2.** *Let  $Q = 1$  be a quadratic equation in a free group  $F_A$ . Then any parametric solution of  $Q = 1$  has a generalization  $\eta$  such that for any variable  $x$ , the length of  $x^\eta$  is bounded by the same function as in Theorem 1.1 with  $c(Q)$  replaced by  $c(Q) + 2n(Q)$ .*

*In particular, there is a finite set of basic parametric solutions of  $Q = 1$  satisfying this bound.*

Note that a simple description of solution sets (in terms of basic parametric solutions) is known for coefficient-free quadratic equations. In this case, only one basic solution is enough. Moreover, if an equation is in the standard form then the value of each variable in the basic parametric solution is either a parameter letter or trivial, see [12, Section 5]. This basic parametric solution is a generalization of every other parametric solution, so the theorem does not give much in this case.

The proof of Theorems 1.1 and 1.2 has three main constituents. In Section 3 we produce various automorphisms that allow to transform quadratic words to a desired form. In

particular, we re-prove a well known fact (closely related to the classification theorem for compact simplicial surfaces, see for example [26, Chapter VI]) that a quadratic word can be equivalently transformed to a word belonging to one of the four standard series, see Proposition 2.1. However, in our proof, we provide an “economical” transformation (Propositions 3.3, 3.5, 3.7, 3.9 and 3.10.) As a corollary we formulate a general result on the bound on the complexity of automorphisms which reduce a given quadratic word to the standard form over *any* group  $G$ .

**Corollary 3.11.** *Let  $Q \in G * F_X$  be quadratic word over an arbitrary group  $G$ . Then there are automorphisms  $\phi, \psi \in \text{Aut}_G(G * F_X)$  such that  $Q^\phi$  and  $Q^\psi$  are conjugate to the standard quadratic words equivalent to  $Q$  and for any variable  $x$ ,*

$$|x^\phi|, |x^{\psi^{-1}}| \leq \begin{cases} 4n(Q) + 2c(Q) & \text{if } Q \text{ is orientable,} \\ 8n^2(Q) + 4n(Q)c(Q) & \text{if } Q \text{ is non-orientable,} \end{cases}$$

where  $c(Q)$  is the total length of coefficients of  $Q$  expressed in any left-invariant (e.g. word) metric on  $G$ .

Note that we provide here bounds for two different automorphisms: for the direct transformation  $Q \xrightarrow{\phi} R$  and for the inverse transformation  $R \xrightarrow{\psi^{-1}} Q$  where  $R$  denotes the standard form of  $Q$ . The reason is that a bound on a automorphism  $\phi$  of a free group  $F$  does not imply a reasonable bound on its inverse  $\phi^{-1}$ ; in particular, see Example 3.8 below.

As the next step of our argument we develop a version of the elimination process for quadratic equations in a free group. To do this, we define a certain set of transformations on pairs of the form  $(Q, \alpha)$  where  $Q$  is a coefficient-free quadratic word and  $\alpha$  is an evaluation of variables in  $Q$ , i.e. an  $F_A$ -homomorphism  $F_{A \cup \text{Var}(Q)} \rightarrow F_A$ . Then we define a certain transformation sequence starting from a given pair  $(Q, \alpha)$  whose primary goal is to eliminate cancellations in the formal word  $Q[\alpha]$  obtained by substituting in  $Q$  of the value  $x^\alpha$  of each variable  $x$ . Our approach is essentially an improved version of similar approaches used in [5, 6, 11] for description of solution sets of quadratic equations in a free group. We start in Section 4 with a lighter version of it, with no care about transformation homomorphisms, and prove Proposition 4.6 sufficient for the proof of Theorem 1.1. Our full version of the eliminated process is elaborated in Section 6 for the proof of Theorem 1.2.

Note that there is an alternating (and essentially equivalent) approach to solutions of quadratic equations in free groups using Lyndon–van Kampen diagrams on surfaces, see for example [24]. Though this approach provides a clear geometric vision, an advantage of using elimination process is an easier control of transformation homomorphisms.

The final ingredient to the proof is a statement about Lyndon–van Kampen diagrams (Proposition 5.4) which says that a diagram can be unfolded in an economical way.

## 2. QUADRATIC WORDS AND SURFACES

In this section, we recall some elementary concepts and formulate some known facts about quadratic words. The free group  $F_A$  and a countable set  $X$  of variables will be fixed throughout the whole paper. The following fact is well known (see for example [12, Section 4]).

**Proposition 2.1.** *By an  $F_A$ -automorphism of  $F_{A \cup X}$ , every quadratic word can be reduced to one of the following forms:*

$$\begin{aligned} & [x_1, y_1][x_2, y_2] \dots [x_g, y_g] \quad (g \geq 0), \\ & x_1^2 x_2^2 \dots x_g^2 \quad (g > 0), \\ & [x_1, y_1][x_2, y_2] \dots [x_g, y_g] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m \quad (g \geq 0, m \geq 1), \\ & x_1^2 x_2^2 \dots x_g^2 c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m \quad (g > 0, m \geq 1). \end{aligned}$$

Here  $x_i, y_i$  and  $z_i$  are variables and  $c_i \in F_A$ ,  $c_i \neq 1$ , are coefficients.

**Definition 2.2.** A quadratic word  $Q$  and an equation  $Q = 1$  are called *standard* if  $Q$  belongs to one of the four series given in Proposition 2.1.

The proof of the proposition essentially repeats the proof of the classification theorem for compact surfaces, see for example a classical topology textbook [26, Sections 38–40]. In the next section we refine the proof and formulate a series of propositions that the reduction automorphism can be chosen economically in terms of its size.

The series of quadratic words in Proposition 2.1 represent the four series of compact surfaces: closed orientable surfaces of genus  $g$ , closed non-orientable surfaces of genus  $g$ , orientable surfaces of genus  $g$  with  $m$  boundary components and non-orientable surfaces of genus  $g$  with  $m$  boundary components, respectively.

The relation between quadratic words and surfaces relies upon the following simple observation. Let  $Q$  be a quadratic word. Take a 2-disk  $D$  with boundary  $\partial D$  divided into  $|Q|$  arcs. Choose an orientation of the boundary  $\partial D$  and label the arcs with letters of  $Q$  in the order they appear in  $Q$  (instead of  $x^{-1}$  we put  $x$  and direct the arc in the opposite to the orientation of  $\partial D$ ). Then for each variable  $x$  occurring in  $Q$ , glue together the two arcs labelled by the two occurrences of  $x$  in  $Q$ , according to their orientation. We get a 2-complex  $S_Q$  representing a compact surface. The surface is closed if  $Q$  is coefficient-free and has a boundary otherwise. (See Fig. 1 where  $Q = x^{-1}ay^{-1}bxcyd$  and  $S_Q$  is homeomorphic to a torus with a disk removed.)

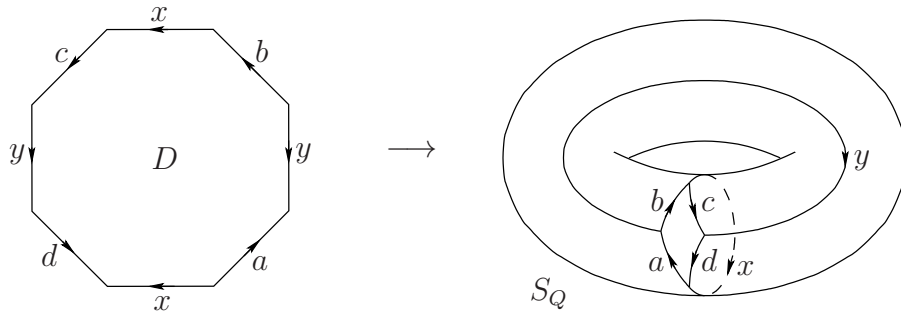


FIGURE 1.

Directed edges of 2-complex  $S_Q$  carry in a natural way labels in  $A^{\pm 1} \cup X^{\pm 1}$ . The edges in the interior of the surface are labelled by variables in  $\text{Var}(Q)$  and edges in the surface boundary labelled by letters in  $A^{\pm 1}$  which come from coefficients of  $Q$ . In case of orientable  $S_Q$  we fix a positive direction of passing along each boundary component which agrees with the chosen orientation of the boundary of disk  $D$ . In this case, we view the label of the boundary

component as an element of  $F_A$  defined up to conjugacy. If  $S_Q$  is non-orientable then labels of boundary components are viewed as elements of  $F_A$  defined up to conjugacy and taking the inverse.

The topological type of  $S_Q$  may be different from the topological type of  $S_R$  where  $R$  is a standard form of  $Q$  from Proposition 2.1. It is not hard to see that if the label  $c$  of a boundary component of  $S_Q$  represents the trivial element of  $F_A$  then the corresponding boundary component disappears in  $S_R$ . We introduce an extra reduction step for  $S_Q$ : if  $c$  is trivial, attach a 2-disk along the corresponding boundary component of  $S_Q$ . We denote  $\bar{S}_Q$  the resulting surface.

**Definition 2.3.** Let  $Q$  be a quadratic word. We call the unordered tuple of all nontrivial labels of boundary components of  $S_Q$ , viewed as elements of  $F_A$ , the *standard coefficients* of  $Q$ . As remarked above, the standard coefficients are defined up to conjugacy if  $S_Q$  is orientable and up to conjugacy and taking inverses if  $S_Q$  is non-orientable.

**Definition 2.4.** Two quadratic words  $Q$  and  $R$  are *equivalent* if  $\bar{S}_Q$  and  $\bar{S}_R$  have the same topological type and their tuples  $(c_{11}, \dots, c_{m1})$  and  $(c_{12}, \dots, c_{m2})$  of standard coefficients are the same up to the natural equivalence: if  $S_Q$  is orientable then up to enumeration, each  $c_{i1}$  is conjugate to the corresponding  $c_{i2}$ ; if  $S_Q$  is non-orientable then up to enumeration, each  $c_{i1}$  is conjugate to either  $c_{i2}$  or  $c_{i2}^{-1}$ .

With a slight abuse of the language, we call a quadratic word  $Q$  itself *orientable* if  $S_Q$  is an orientable surface and *non-orientable* otherwise. It is immediate from the construction of  $S_Q$  that  $Q$  is orientable if and only if each variable in  $Q$  has two occurrences with different exponents  $-1$  and  $+1$ . By the *genus* of  $Q$  we mean the genus of  $S_Q$ .

Thus, any quadratic word  $Q$  is defined up to equivalence by the following parameters: orientability (true/false), the genus  $g$  and the tuple of standard coefficients  $(c_1, \dots, c_m)$ . Proposition 2.1 asserts precisely that any quadratic word can be reduced by an  $F_A$ -automorphism of  $F_{A \cup X}$  to an equivalent standard quadratic word.

The following fact may be attributed to folklore. (We don't make use of it in the paper and provide only a sketch of the proof.)

**Proposition 2.5.** *Let  $Q, R \in F_{A \cup X}$  be two quadratic words. Then the following statements are equivalent:*

- (i)  $R$  is equivalent to  $Q$ ;
- (ii)  $R$  is the image of  $Q$  under an  $F_A$ -automorphism of  $F_{A \cup X}$ ;
- (iii)  $R$  is conjugate to the image of  $Q$  under an  $F_A$ -automorphism of  $F_{A \cup X}$ .

*Sketch of the proof:* To prove implication (i) $\Rightarrow$ (ii), we reduce two given equivalent quadratic words to equivalent standard forms by Proposition 2.1. It is then an easy exercise to show that two equivalent standard quadratic words are images of each other under an  $F_A$ -automorphisms of  $F_{A \cup X}$  (one can use also automorphisms from Lemmas 3.13 and 3.14 below).

Implication (iii) $\Rightarrow$ (i) follows easily from the following statement: If  $Q, R \in F_{A \cup X}$  are quadratic words and  $R$  is conjugate to  $Q^\phi$  for some  $\phi \in \text{Aut}_{F_A}(F_{A \cup X})$  then  $\phi$  may be represented as a product  $\tau_1 \tau_2 \dots \tau_k$  of elementary Nielsen  $F_A$ -automorphisms  $\tau_i$  such that the cyclically reduced form of  $Q^{\tau_1 \dots \tau_i}$  is a quadratic word for each  $i$ . To prove the statement, we use the Higgins–Lyndon approach to stabilizers in  $\text{Aut}(F)$  as exposed in Lyndon and Schupp's book [21, Section I.4]. We apply a modified version of Proposition I.4.23 of [21] to



one cyclic word  $u_1 = Q$  and the tuple of non-cyclic words  $u_i$ ,  $i \geq 2$ , consisting of all letters  $a \in A$ . By this proposition,  $\phi$  can be represented as a product

$$\phi = \rho_1 \rho_2 \dots \rho_r$$

of Whitehead automorphisms  $\rho_i$  such that for some  $p$  and  $q$ ,  $0 \leq p \leq q \leq r-1$ ,

$$|Q_{i+1}| < |Q_i| \text{ for } i = 0, \dots, p, \quad |Q_{i+1}| = |Q_i| \text{ for } i = p+1, \dots, q$$

and

$$|Q_{i+1}| > |Q_i| \text{ for } i = q+1, \dots, r-1$$

where, by definition,  $Q_0 = Q$  and  $Q_i$  is the cyclically reduced form of  $Q^{\rho_1 \rho_2 \dots \rho_i}$ . Then it is not hard to see that in each of the chains

$$Q_0 \xrightarrow{\rho_1} Q_1 \xrightarrow{\rho_2} \dots \xrightarrow{\rho_q} Q_q \quad \text{and} \quad Q_r \xrightarrow{\rho_r^{-1}} Q_{r-1} \xrightarrow{\rho_{r-1}^{-1}} \dots \xrightarrow{\rho_{q+1}^{-1}} Q_q$$

every automorphism  $\rho_i^{\pm 1}$  can be factored into a sequence of elementary Nielsen automorphisms which keep the property of a word being quadratic.  $\square$

### 3. TRANSFORMING QUADRATIC WORDS

In this section, we describe several specific transformations of quadratic words to produce equivalent quadratic words of a desired form. By a transformation we mean application of an  $F_A$ -automorphism of  $F_{A \cup X}$  (or an  $F_C$ -automorphism of  $F_{C \cup X}$  where  $F_C$  is the group of formal coefficients, see below).

By  $|W|$  we denote the length of an element of a free group written as a freely reduced word. The notation  $|W|_x$  will be used for the total number of occurrences of a letter  $x$  in  $W$ . More generally, if  $S$  is a set of letters then  $|W|_S$  will denote the total number of occurrences in  $W$  of letters from  $S$ . Sometimes we consider formal (not necessarily freely reduced) words  $W$ . In this case,  $|W|$  denotes the length of  $W$ .

For automorphisms of a free group  $F_Y$ , we use a notation

$$\phi = (x_1^{\varepsilon_1} \mapsto W_1, x_2^{\varepsilon_2} \mapsto W_2, \dots, x_k^{\varepsilon_k} \mapsto W_k) \quad \text{where } x_i \in Y, \varepsilon_i = \pm 1, W_i \in F_Y$$

which means that  $\phi$  maps generator  $x_i^{\varepsilon_i}$  to  $W_i$  and does not change other generators.

We describe several types of elementary transformations  $\phi$  applied to a quadratic word  $Q \in F_{A \cup X}$  and producing an equivalent quadratic word  $Q^\phi$ , or, in a weaker form, an equivalent quadratic word conjugate to the image  $Q^\phi$  of  $Q$ .

*Permutations and exponent sign changes of variables:* Permutations on the set of variables and automorphisms of the form  $(x \mapsto x^{-1})$ . These automorphisms always carry quadratic words into equivalent ones. We will implicitly assume that automorphisms of this type are applied if needed. For example, to transform a word  $Q_1$  to a given word  $Q_2$  it is enough to find an automorphism  $\phi \in \text{Aut}_{F_A}(F_{A \cup X})$  such that  $Q_1^\phi$  is equal to  $Q_2$  up to renaming variables and changing their exponent signs.

*Multiplications by coefficients:* Automorphisms of the form  $(x^\varepsilon \mapsto x^\varepsilon d)$ ,  $d \in F_A$ . Any automorphism of this form carries any quadratic word to an equivalent one. We will use only the special case when  $x^\varepsilon d^{-1}$  or  $dx^{-\varepsilon}$  occurs in  $Q$  as a subword. We refer to these transformations as *coefficient shifts*. Geometrically, we shift the start of the edge labelled  $x^\varepsilon$  along the boundary arc of  $S_Q$  labelled  $d^{-1}$  from the start of this arc to its end (Fig. 2). This automorphism does not change  $S_Q$  unless the corresponding boundary component is labelled by the trivial element and there are two occurrences of  $(x^\varepsilon d^{-1})^{\pm 1}$  in  $Q$  (Fig. 3). In this case,

this boundary component disappears (but the topological type of the reduced surface  $\bar{S}_Q$  is not changed).

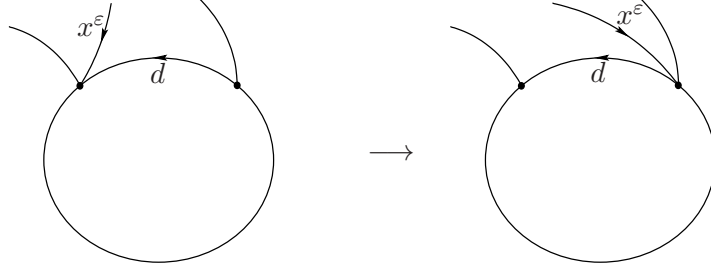


FIGURE 2.

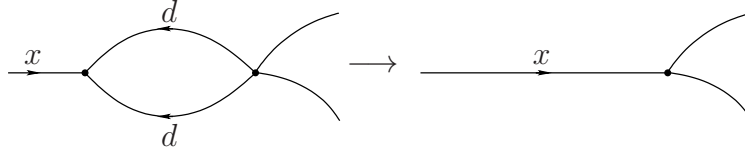


FIGURE 3.

*Introductions of new variables:* Nielsen automorphisms of the form  $(x^\varepsilon \mapsto x^\varepsilon y^\delta)$  where a variable  $y$  does not occur in  $Q$ .

*Related Nielsen automorphisms:* Nielsen automorphisms  $(x^\varepsilon \mapsto x^\varepsilon y^\delta)$  related to  $Q$ , that is, those for which  $(x^\varepsilon y^{-\delta})^{\pm 1}$  occurs in  $Q$ . Geometrically, we shift the start of the edge labelled  $x^\varepsilon$  along the edge labelled  $y^\delta$ , see Fig. 4. If  $(x^\varepsilon y^{-\delta})^{\pm 1}$  occurs in  $Q^{\pm 1}$  twice then  $y$  is eliminated from  $Q$ . If we view  $Q$  as a cyclic word then for any Nielsen automorphism  $\rho$  related to  $Q$ , the image  $Q^\rho$  is conjugate to a quadratic word equivalent to  $Q$ .

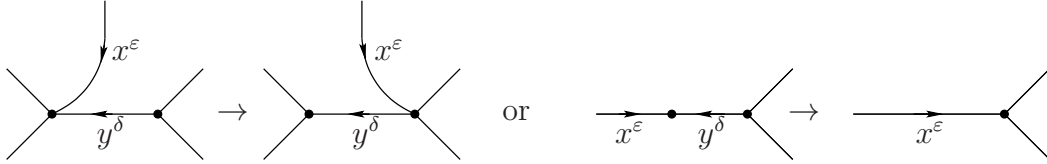


FIGURE 4.

It is sometimes convenient to view transformations of quadratic words from a slightly different angle—through formal coefficients. This means that we introduce a new alphabet  $C$  of *formal coefficients* and consider quadratic words in  $Q \in F_{C \cup X}$  with the property that each coefficient letter  $c \in C$  occurs in  $Q$  at most once. To get a quadratic word in the usual sense (as an element of  $F_{A \cup X}$ ) we have to provide it with a *coefficient map*  $F_C \rightarrow F_A$ .

Thus we have two ways of representing quadratic words: as an element of  $F_{A \cup X}$  and as a pair  $(Q, \gamma)$  where  $Q$  is a quadratic word with formal coefficients and  $\gamma$  is a coefficient map. An advantage of the second way is that quadratic words can be viewed independently on the coefficient group which may be not necessarily free. We use this representation since it provides a more convenient accounting of lengths of coefficients occurring in transformations.



Note that all elementary transformations introduced above are applicable also to quadratic words with formal coefficients (just with  $F_A$  replaced by the formal coefficient group  $F_C$ ). If a pair  $(Q, \gamma)$  represents a quadratic word  $R \in F_{A \cup X}$  then transformations of  $Q$  induce corresponding transformations of  $R$ .

In the rest of the section we assume that quadratic words are ones with formal coefficients and belong to the group  $F_{C \cup X}$ .

We prove a series of statements asserting that a quadratic word can be reduced to a standard form using an automorphism of bounded complexity. Note that the notion of a standard quadratic word is essentially not changed when passing to words with formal coefficients. It is required only that any formal coefficient occurs in the coefficients  $c_i$  of a standard quadratic word at most once in total.

**Definition 3.1.** We call automorphisms of the form  $(x^\varepsilon \rightarrow Wx^\varepsilon)$  where  $x$  does not occur in  $W$ , *elementary*.

We say that an elementary automorphism  $(x^\varepsilon \mapsto Wx^\varepsilon)$  *changes*  $x$  and *touches* variables occurring in  $W$ .

A product  $\rho_1 \rho_2 \dots \rho_k$  of elementary automorphisms  $\rho_i$  is *triangular* if it satisfies the following condition: as soon as  $\rho_i$  touches  $x$ , the variable  $x$  is not changed by all subsequent automorphisms  $\rho_{i+1}, \dots, \rho_k$ .

The following observation is immediate:

**Lemma 3.2.** *Let*

$$\phi = (x_1^{\varepsilon_1} \rightarrow W_1 x_1^{\varepsilon_1})(x_2^{\varepsilon_2} \rightarrow W_2 x_2^{\varepsilon_2}) \dots (x_k^{\varepsilon_k} \rightarrow W_k x_k^{\varepsilon_k})$$

*be a triangular product of elementary automorphisms of a free group  $F$ . For a generator  $x$ , let  $W_{i_1}, W_{i_2}, \dots, W_{i_r}$  be all words  $W_i$  which participate in automorphisms  $(x_i \rightarrow W_i x_i)$  with  $x_i = x$  and  $\varepsilon_i = 1$ , and  $W_{j_1}, W_{j_2}, \dots, W_{j_t}$  be all words  $W_i$  which participate in automorphisms  $(x_i^{-1} \rightarrow W_i x_i^{-1})$  with  $x_i = x$  and  $\varepsilon_i = -1$ . Then*

$$x^\phi = W_{i_1} W_{i_2} \dots W_{i_r} x W_{j_t}^{-1} W_{j_{t-1}}^{-1} \dots W_{j_1}^{-1}.$$

*In particular, for any generator  $y \neq x$ , the number of occurrences of  $y$  in  $x^\phi$  does not exceed the total number of occurrences of  $y$  in all  $W_{i_k}$  and  $W_{j_k}$ .*

**Proposition 3.3.** *Let  $Q \in F_{C \cup X}$  be an orientable quadratic word. Then there exists an automorphism  $\phi \in \text{Aut}_{F_C}(F_{C \cup \text{Var}(Q)})$  such that  $Q^\phi$  is conjugate to a standard quadratic word equivalent to  $Q$  and for any  $x, y \in \text{Var}(Q)$  and  $c \in C$ , we have  $|x^\phi|_y \leq 4$  and  $|x^\phi|_c \leq 2$ .*

*In particular,  $|x^\phi| \leq 2|Q|$  for any  $x \in \text{Var}(Q)$ ,*

*Proof.* We construct the required automorphism  $\phi$  as a triangular product of elementary automorphisms. Starting from this point throughout the proof, we denote by  $Q$  the current quadratic word after application of a sequence of elementary automorphisms constructed so far. At the start,  $Q$  is any quadratic word from the hypothesis of the proposition. We view  $Q$  as a cyclic word and thus regard transformations up to conjugation.

We assume that  $Q$  has at least one variable (otherwise the proposition is trivial).

At any moment, there are *locked* variables in  $Q$  which have been touched by previous elementary automorphisms. They should not be changed by subsequent elementary automorphisms.

By  $\Gamma_Q$  we denote the 1-skeleton of  $S_Q$ , i.e. the graph embedded in the surface after the identification of arcs in the boundary of disk  $D$  as described in Section 2.

*Step 1: Eliminating boundary superfluous vertices.*

If  $S_Q$  is a closed surface then Step 1 is void and we jump to Step 2. Recall that  $S_Q$  is closed if and only if  $Q$  is coefficient-free.

We call a vertex  $\nu$  in the boundary of  $S_Q$  *essential* if it is an endpoint of an interior edge of  $S_Q$  (i.e. an endpoint of an edge labelled by a variable).

Let  $\ell$  be a boundary component of  $S_Q$ . Observe that  $\ell$  has at least one essential vertex (since  $Q$  has at least one variable). We describe a sequence of elementary automorphisms which result in exactly one essential vertex in  $\ell$ . Let  $\nu_1, \dots, \nu_k$  be all cyclically ordered essential vertices in  $\ell$ . Let  $c \in F_A$  be the label of the arc between  $\nu_1$  and  $\nu_2$  and  $x_1, \dots, x_r \in X^{\pm 1}$  be labels of edges starting at  $\nu_2$  so that  $cx_1, x_1^{-1}x_2, \dots, x_{r-1}^{-1}x_r$  occur in  $Q$  (see Fig. 5). We apply to  $Q$  a sequence of elementary automorphisms

$$(x_1 \mapsto c^{-1}x_1)(x_2 \mapsto c^{-1}x_2) \dots (x_r \mapsto c^{-1}x_r)$$

eliminating essential vertex  $\nu_2$ . Then we proceed in the same way eliminating all other essential vertices  $\nu_3, \dots, \nu_k$ .

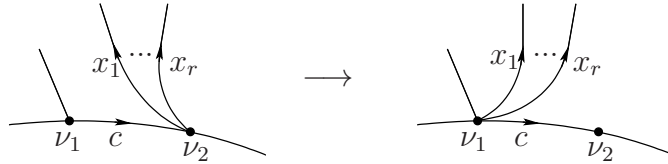


FIGURE 5.

We repeat the procedure for all other boundary components of  $S_Q$ . After that, each boundary component will have exactly one essential vertex. Their labels are the standard coefficients of  $Q$ . There are no locked variables after this step.

*Step 2: Eliminating inner superfluous vertices.*

We choose a base vertex  $\nu_1$  of  $\Gamma_Q$  as follows. If  $Q$  is coefficient-free we take any vertex of  $\Gamma_Q$ . If  $Q$  has a coefficient then for the base vertex we take any essential vertex in the boundary of  $S_Q$ .

Suppose that  $\Gamma_Q$  has a vertex  $\nu \neq \nu_1$  in the interior of  $S_Q$ . Let  $e_1, \dots, e_k$  be all directed edges starting at  $\nu$ , labelled by variables  $x_1, \dots, x_k \in X^{\pm 1}$ . Since  $\nu$  and  $\nu_1$  are connected by a path in  $\Gamma_Q$ , at least one of  $e_i$ , say  $e_1$ , ends in a vertex  $\nu'$  distinct from  $\nu$ . Then we apply a sequence of elementary automorphisms

$$\psi = (x_2 \mapsto x_1^{-1}x_2) \dots (x_k \mapsto x_1^{-1}x_k)$$

eliminating  $\nu$ . Observe that  $x_1$  does not occur in the new word  $Q^\psi$ , so the locked variable  $x_1$  will not participate in the subsequent automorphisms.

We perform elimination of all vertices  $\nu \neq \nu_1$  in the interior of  $S_Q$ . After that, if  $Q$  is coefficient-free then  $\Gamma_Q$  has only one vertex  $\nu_1$ . If  $Q$  has a coefficient then  $\Gamma_Q$  has  $m$  vertices  $\nu_1, \dots, \nu_m$ , one in each boundary component of  $S_Q$ .

*Step 3: Collecting coefficient factors.*

We assume here that  $Q$  has at least one coefficient. If  $Q$  is coefficient-free, we jump to Step 4.

The step consists of a sequence of substeps  $3_1, 3_2, \dots, 3_{m-1}$ . Before step  $3_i$ ,  $Q$  has the form

$$Q = c_1 z_2^{-1} c_2 z_2 z_3^{-1} c_3 z_3 \dots z_i^{-1} c_i z_i W$$

where  $z_j \in X^{\pm 1}$  and  $c_j \in F_A$  and  $W$  has no locked variables. Recall that  $Q$  is viewed as a cyclic word, so at the start of step 3 we have  $Q = c_1 W$  for some  $W$  and coefficient  $c_1$ .

If no coefficients occur in  $W$  then we stop. Suppose that a coefficient occurs in  $W$ . Since  $\Gamma_Q$  is connected, there is an edge in  $\Gamma_Q$  starting at  $\nu_1$  and ending in a boundary component of  $S_Q$  distinct from ones labelled by  $c_1, c_2, \dots, c_i$ . Let  $z_{i+1} \in X^{\pm 1}$  be such an edge (for convenience we identify edges with their labels) and  $\nu_{i+1}$  its endpoint in a boundary component labelled  $c_{i+1}$ .

Step  $3_i$  consists of the following. Using related Nielsen automorphisms and coefficient shifts, we first shift starting vertices of all interior edges of  $S_Q$  at  $\nu_{i+1}$  other than  $z_{i+1}^{-1}$  to a new position at  $\nu_1$  along the path labelled  $z_{i+1}^{-1}$  or the path labelled  $c_{i+1} z_{i+1}^{-1}$  (see Fig. 6). Next, if there are edges starting at  $\nu_1$  between  $z_i$  and  $z_{i+1}$ , we shift them one by one along the path labelled  $z_{i+1} c_{i+1} z_{i+1}^{-1}$ .

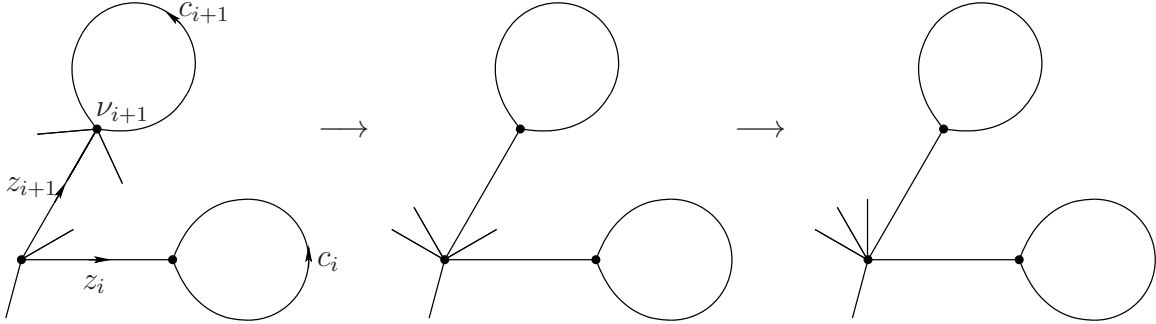


FIGURE 6.

After that,  $Q$  gets the form

$$Q = c_1 z_2^{-1} c_2 z_2 z_3^{-1} c_3 z_3 \dots z_{i+1}^{-1} c_{i+1} z_{i+1} W'$$

and we iterate the procedure. Finally we come to a word of the form

$$Q = c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m R$$

where  $R$  is a coefficient-free quadratic word with no locked variables. If  $R$  is empty we get the desired standard quadratic word. Otherwise we proceed to the next step 4.

Observe that all edges of  $\Gamma_Q$  labelled by variables occurring in  $R$  start and end at the same vertex  $\nu_1$ . In this case, we define a *star word*  $R^*$  as the sequence of labels of edges labelled by variables in  $R$  when moving around  $\nu_1$  in a small neighborhood of  $\nu_1$ . To fix the direction of the motion we agree that if  $x^\varepsilon y^\delta$  occurs in  $R^*$  then  $y^{-\delta} x^\varepsilon$  occurs in  $R$ . If  $Q$  has a coefficient then  $R^*$  is the word read off between the edges of  $\Gamma_Q$  labelled the starting letter of  $c_1$  and  $z_m$  (see Fig. 7). If  $Q$  is coefficient-free then  $R = Q$  and we view  $R^*$  as a cyclic word.

Observe that  $R^*$  is an orientable quadratic word of the same length and with the same variables as  $R$ .

*Step 4: Collecting commutators.* We assume that  $Q$  has a coefficient; the coefficient-free case is similar with obvious minor changes.

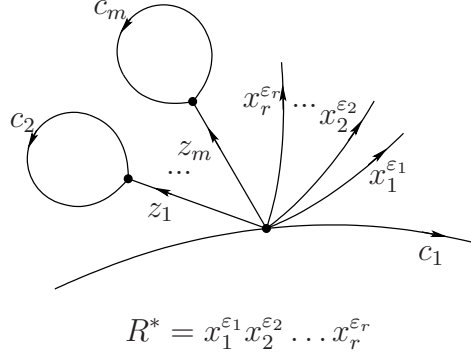


FIGURE 7.

At any stage of this step, we do not change the coefficient part of  $Q$  already collected, so  $Q$  has the above form

$$Q = c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m R$$

where  $R$  is a coefficient-free quadratic word. We will apply only related Nielsen automorphisms involving variables of  $R$ . This implies that all edges of  $\Gamma_Q$  labelled by variables in  $R$  start and end at the same vertex  $\nu_1$ .

The whole step 4 is again an iterative sequence of smaller steps  $4_1, \dots, 4_g$ . Before substep  $4_i$ , we have

$$R = [x_1, y_1] \dots [x_{i-1}, y_{i-1}] T$$

and

$$R^* = T^* [y_{i-1}, x_{i-1}^{-1}] \dots [y_1, x_1^{-1}]$$

where  $T$  and  $T^*$  are orientable quadratic words with  $\text{Var}(T) = \text{Var}(T^*)$ ; in particular,  $|T| = |T^*|$ .

We stop if  $T$  and  $T^*$  are empty.

Assume that  $T$  and  $T^*$  are nonempty. Let  $x_i$  be a variable occurring in  $T$ , so

$$T^* = U x_i V x_i^{-1} W$$

up to the exponent sign of  $x_i$ .

We claim that at least one variable  $y_i$  occurs in  $V$  exactly once. Indeed, if  $V$  is a (possibly empty) quadratic word then a quadratic word  $\bar{V}$  lies between the two occurrences of  $x_i$  in  $T$ , that is,  $T = \dots x_i^{\pm 1} \bar{V} x_i^{\mp 1} \dots$ . In this case the edge of  $\Gamma_Q$  labelled  $x_i$  would have distinct endpoints, a contradiction.

Hence, up to interchanging  $x_i$  and  $y_i$  and changing their exponent signs we have

$$T^* = Z_1 y_i^{-1} Z_2 x_i Z_3 y_i Z_4 x_i^{-1} Z_5$$

The following sequence of automorphisms collects the commutator  $[y_i, x_i^{-1}]$  when applied to  $T^*$ :

$$\begin{aligned}
T^* &\xrightarrow{(x_i \mapsto Z_5 x_i)} Z_1 y_i^{-1} Z_2 Z_5 x_i Z_3 y_i Z_4 x_i^{-1} \\
&\xrightarrow{(y_i \mapsto Z_2 Z_5 u_i)} Z_1 y_i^{-1} x_i Z_3 Z_2 Z_5 y_i Z_4 x_i^{-1} \\
&\xrightarrow{(x_i \mapsto x_i (Z_3 Z_3 Z_5)^{-1})} Z_1 y_i^{-1} x_i y_i Z_4 Z_3 Z_2 Z_5 x_i^{-1} \\
&\xrightarrow{(y_i \mapsto y_i (Z_4 Z_3 Z_3 Z_5)^{-1})} Z_1 Z_4 Z_3 Z_2 Z_5 y_i^{-1} x_i y_i x_i^{-1}
\end{aligned}$$

It is straightforward to check that for any Nielsen automorphism  $\rho$  related to  $T^*$  there is a dual Nielsen automorphism  $\rho^*$  related to  $T$  whose action on  $T$  agrees with the action of  $\rho$  on  $T^*$ , that is,  $(T^*)^\rho = (T^{\rho^*})^*$ : If  $x^\varepsilon y^\delta$  occurs in  $T^*$  then we define

$$(x^\varepsilon \mapsto x^\varepsilon y^{-\delta})^* = (y^\delta \mapsto x^\varepsilon y^\delta) \quad \text{and} \quad (y^\delta \mapsto x^{-\varepsilon} u^\delta)^* = (x^\varepsilon \mapsto y^\delta x^\varepsilon).$$

We observe that if  $\rho$  changes  $x$  and touches  $y$ , then the role of these variables in  $\rho^*$  is interchanged. This implies that there is a sequence  $\psi$  of related Nielsen automorphisms touching only  $x_i$  and  $y_i$  which transforms  $T$  to a word  $T^\psi$  where

$$(T^\psi)^* = Z[y_i, x_i^{-1}].$$

Then for some  $T'$ ,

$$T^\psi = [x_i, y_i] T'$$

as required. This finishes step 4<sub>i</sub>.

After completion of this procedure  $Q$  gets the standard form. We get an automorphism  $\phi$  reducing to such a form an arbitrary quadratic word from the hypothesis of the proposition.

*Calculation of the bounds for  $\phi$ .* Fix a variable  $x \in \text{Var}(Q)$ . Let  $W_1, W_2, \dots, W_k$  be the list of all words  $W$  in elementary automorphisms  $(x^\varepsilon \rightarrow Wx^\varepsilon)$  which are the factors of  $\phi$ . We claim that any variable  $y \neq x$  occurs in all  $W_i$  in total at most 4 times. Indeed, at each individual step 2, 3 or 4,  $y$  occurs at most twice in the automorphisms of the form  $(x \rightarrow Wx)$  and at most twice in the automorphisms of the form  $(x^{-1} \rightarrow Wx^{-1})$ . It remains to observe that if  $y$  participates (i.e. occurs in some  $W$ ) at step 2 or 3 then  $y$  does not participate in subsequent steps. By Lemma 3.2 we get

$$|x^\phi|_y \leq 4.$$

From the construction it is easy to see also that each constant  $c \in C$  occurs in  $|x^\phi|$  at most twice. This implies

$$|x^\phi|_c \leq 2.$$

This finishes the proof of Proposition 3.3. □

**Definition 3.4.** We call a non-orientable quadratic word  $Q$  *semi-standard* if  $Q$  has one of the following forms

$$Q = x_1^2 x_2^2 \dots x_k^2 [x_{k+1}, y_{k+1}] \dots [x_n, y_n]$$

or

$$Q = x_1^2 x_2^2 \dots x_k^2 [x_{k+1}, y_{k+1}] \dots [x_n, y_n] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m$$

where factors  $[x_i, y_i]$  are not obligatory.

**Proposition 3.5.** *Let  $Q \in F_{C \cup X}$  be a non-orientable quadratic word. Then there exists an automorphism  $\phi \in \text{Aut}_{F_C}(F_{C \cup \text{Var}(Q)})$  such that  $Q^\phi$  is conjugate to a semi-standard quadratic word equivalent to  $Q$  and for any  $x, y \in \text{Var}(Q)$  and  $c \in C$ , we have  $|x^\phi|_y \leq 4$  and  $|x^\phi|_c \leq 2$ . In particular,  $|x^\phi| \leq 2|Q|$  for any  $x \in \text{Var}(Q)$ .*

*Proof.* The steps 1–3 are the same as in the case of orientable  $S_Q$ . After performing these three steps we get

$$Q = c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m R$$

where  $R$  is now a non-orientable coefficient-free quadratic word. The rest of reduction consists of the following step.

*Step 4<sup>n</sup>:* Collecting squares and commutators.

As in the orientable case, we will work with the star word  $R^*$ . Unfortunately, the definition of  $R^*$  given in the proof of Proposition 3.3 suits only for orientable  $R$  since it always produces an orientable  $R^*$ . We modify the definition.

For any edge  $e$  of  $S_Q$ , we fix its *orientation* which is a choice of the positive direction of crossing  $e$  inside  $S_Q$ . When passing around the base vertex  $\nu_1$ , we read the label  $x$  of  $e$  with exponent  $+1$  if we cross  $e$  in the positive direction and  $-1$  otherwise (so the exponent signs of variables in  $R^*$  are not related directly to the exponent signs of the corresponding occurrences in  $R$ ). It is easy to check that application of a related Nielsen automorphism  $\rho = (x^\varepsilon \mapsto y^\delta x^\varepsilon)$  to  $R^*$  agrees with application of a dual Nielsen automorphism  $\rho^*$  related to  $R$  that changes  $y$  and touches  $x$ .

Let  $e$  be an edge of  $\Gamma_Q$  labelled by a variable  $x \in \text{Var}(R)$ . If  $e$  reverses the orientation when viewed as a loop in  $S_Q$  then we cross  $e$  twice in the same direction when passing around  $\nu_1$ . Hence the both occurrences of  $x$  in  $R^*$  have the same exponent sign. Since there is at least one orientation-reversing  $e$ ,  $R^*$  is a non-orientable quadratic word. Thus, up to the exponent sign of  $x$ ,  $R^*$  has the form

$$R^* = Z_1 x Z_2 x Z_3.$$

We then collect the square of  $x$ :

$$R^* \xrightarrow{(x \mapsto x Z_3^{-1})} Z_1 x Z_3^{-1} Z_2 x \xrightarrow{(x \mapsto Z_2^{-1} Z_3 x)} Z_1 Z_2^{-1} Z_3 x^2$$

If  $Z_1 Z_2^{-1} Z_3$  is non-orientable then we repeat the procedure collecting the square of a variable to the right. Otherwise we apply the procedure of collecting commutators described in Step 4 in the orientable case. Using a triangular sequence of elementary automorphisms we finally reduce  $Q$  to the form

$$Q = c_0 z_1^{-1} c_1 z_1 \dots z_{m-1}^{-1} c_{m-1} z_{m-1} x_1^2 \dots x_k^2 [x_{k+1}, y_{k+1}] \dots [x_r, y_r]$$

The bound on  $\phi$  is obtained in the same way as in the orientable case. □

To reduce a semi-standard non-orientable quadratic word to a standard form we need an extra transformation.

**Lemma 3.6.** *There are automorphisms  $\eta_k, \theta_k \in \text{Aut}(F_{\{x_0, \dots, x_k, y_1, \dots, y_k\}})$  such that*

$$x_0^2 [x_1, y_1] \dots [x_k, y_k] \xrightarrow{\eta_k, \theta_k} x_0^2 x_1^2 y_1^2 \dots x_k^2 y_k^2$$

*and for any  $x \in \{x_0, \dots, x_k, y_1, \dots, y_k\}$ ,*

$$|x^{\eta_k}| \leq 4k + 1 \quad \text{and} \quad |x^{\theta_k^{-1}}| \leq 3k + 2.$$



*Proof.* Let  $\gamma(x, y, z)$  be an automorphism of  $F_{\{x, y, z\}}$  such that

$$(x^2[y, z])^\gamma = x^2 y^2 z^2.$$

Then taking

$$\eta_k = \gamma(x_0, x_1, y_1) \gamma(y_1, x_2, y_2) \dots \gamma(y_{k-1}, x_k, y_k)$$

we obviously get

$$x_0^2[x_1, y_1] \dots [x_k, y_k] \xrightarrow{\eta_k} x_0^2 x_1^2 y_1^2 \dots x_k^2 y_k^2$$

For a specific  $\gamma$ , we take

$$\gamma = (x \mapsto x^2 y z x^{-1}, y \mapsto x y z x^{-1}, z \mapsto x z).$$

The bound  $\|\eta_k\| \leq 4k + 1$  is straightforward.

To define  $\theta_k$  we proceed in a similar way by taking for  $\gamma$  another automorphism

$$\gamma = (x \mapsto x y z, y \mapsto z^{-1} y^{-1} x^{-1} y z x y z, z \mapsto z^{-1} y^{-1} x^{-1} z)$$

with

$$\gamma^{-1} = (x \mapsto x^2 y^{-1} x^{-1}, y \mapsto x y x^{-1} z^{-1} x^{-1}, z \mapsto x z).$$

□

**Proposition 3.7.** *Let  $Q \in F_{C \cup X}$  be a non-orientable quadratic word. Then there exists an automorphism  $\phi \in \text{Aut}_{F_C}(F_{C \cup \text{Var}(Q)})$  such that  $Q^\phi$  is conjugate to a standard quadratic word equivalent to  $Q$  and for any  $x \in \text{Var}(Q)$  we have  $|x^\phi|_X \leq 8n(Q) \text{genus}(Q)$  and  $|x^\phi|_c \leq 2$  for any formal coefficient  $c \in C$ .*

*Proof.* Proposition 3.5 and Lemma 3.6 produce an automorphism  $\phi \in \text{Aut}_{F_C}(F_{C \cup \text{Var}(Q)})$  such that  $Q^\phi$  is conjugate to a standard quadratic word equivalent to  $Q$  and

$$|x^\phi|_X \leq 4n(Q)(4k + 1) \quad \text{and} \quad |x^\phi|_c \leq 2$$

where  $k$  is the number of commutators in the semi-standard form given by Proposition 3.5. It remains to notice that

$$k \leq \frac{1}{2}(\text{genus}(Q) - 1).$$

□

We turn now to bounds similar to Propositions 3.3, 3.5 and 3.7 where we estimate the size of the automorphism *inverse* to  $\phi$ . Note that the sizes of an automorphism  $\phi$  of a free group  $F$  and of its inverse  $\phi^{-1}$  can be very different. We give an example where the ratio is exponential in the rank of  $F$ .

**Example 3.8.** Let  $F = F_{\{x_1, \dots, x_n\}}$ . Define an automorphism  $\phi \in \text{Aut}(F)$  by

$$\begin{aligned} x_{2i+1}^\phi &= x_{i+1} x_i x_{i+2} x_{i-1} \dots x_{2i} x_1 x_{2i+1}, \\ x_{2i}^\phi &= x_i x_{i+1} x_{i-1} x_{i+2} \dots x_1 x_{2i} \end{aligned}$$

Then  $\|\phi\| = n$  where, by definition,  $\|\phi\| = \max_i |x_i^\phi|$ . It is not hard to see that  $\|\phi^{-1}\| = 2^n$ .

**Proposition 3.9.** *Let  $Q \in F_{C \cup X}$  be any quadratic word. Then there exists an automorphism  $\psi \in \text{Aut}_{F_C}(F_{C \cup \text{Var}(Q)})$  such that  $Q^\psi$  is conjugate to a quadratic word  $R$  equivalent to  $Q$  and the following assertions are true:*

- (i)  $R$  is standard if  $Q$  is orientable and semi-standard if  $Q$  is non-orientable.

- (ii) For any  $x, y \in \text{Var}(Q)$  and  $c \in C$ , we have  $|x^{\psi^{-1}}|_y \leq 4$  and  $|x^{\psi^{-1}}|_c \leq 2$ . In particular,  $|x^{\psi^{-1}}| \leq 2|Q|$ .

*Proof.* Similarly to the arguments used in the proof of Propositions 3.3 and 3.5, we view  $Q$  as a cyclic word and construct  $\psi$  so that the inverse automorphism  $\psi^{-1}$  will be a triangular product of elementary automorphisms. The condition that  $\psi^{-1}$  is triangular is equivalent to the condition that  $\psi$  is *reverse triangular* in the following sense: if a variable  $x$  is changed by some elementary automorphism in the product then  $x$  is not touched by subsequent elementary automorphisms. In fact, the proof will be simpler than the proof of Propositions 3.3 and 3.5 because there is no need to pass to the star word and we will operate on the quadratic word  $Q$  itself.

According to the change in the notion of a triangular product, we change the notion of a locked variable: a variable  $x$  is viewed as locked at a current transformation step if it was previously changed by an elementary automorphism. Subsequent elementary automorphisms should not touch locked variables. Similar to the proof of Propositions 3.3 and 3.5, we collect locked variables to a fixed part of  $Q$ . Thus, at any moment  $Q$  has the form  $LR$  where  $L$  is the locked part of  $Q$  and  $R$  has no locked variables. At the start,  $L$  is empty and since  $Q$  is viewed as a cyclic word, we assume without loss of generality that  $R$  starts with a coefficient letter.

We proceed in two steps.

*Step 1<sup>i</sup>:* Reducing the coefficient-free part.

If  $R$  is non-orientable then  $R = Z_1 x Z_2 x Z_3$  for some variable  $x$ . Using a sequence of elementary automorphisms similar to one given in Step 4<sup>n</sup> in the proof of Proposition 3.5 we collect  $x^2$  to the left of  $R$  and add it to the locked part.

If  $R$  is orientable and there are variables  $x$  and  $y$  which “cross” in  $R$ , that is,  $R = Z_1 x^{-1} Z_2 y^{-1} Z_3 x Z_4 y Z_5$  (up to exponent signs of  $x$  and  $y$ ) then we collect the commutator  $[x, y]$  to the left of  $R$  as in Step 4 in the proof of Proposition 3.3.

Iterating the procedure, we reduce  $Q$  to the form

$$Q = x_1^2 \dots x_k^2 [x_{k+1}, y_{k+1}] \dots [x_r, y_r] R$$

where  $R$  is orientable and has no crossing variables.

*Step 2<sup>i</sup>:* Collecting the coefficient part.

If  $R$  has no variables then we are done. Let  $\text{Var}(R) \neq \emptyset$ . By the assumption that  $R$  has no crossing variables, there is a variable  $z$  with no variables between the two occurrences of  $z$  in  $R$ . Then we have  $R = T_1 z^{-1} c z T_2$  where  $c \in F_A$  is a standard coefficient of  $Q$ . Applying automorphism  $(z \mapsto z T_2^{-1})$  to  $Q$  we shift  $z^{-1} c z$  to the right of  $R$ . Iterating the procedure we get

$$Q = x_1^2 \dots x_k^2 [x_{k+1}, y_{k+1}] \dots [x_r, y_r] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m$$

It remains to observe that at any transformation step, we preserve the property that  $R$  starts with a coefficient letter and so  $c_1 \neq 1$ .

The required bounds  $|x^{\psi^{-1}}|_y \leq 4$  and  $|x^{\psi^{-1}}|_c \leq 2$  are straightforward in view of Lemma 3.2 (applied to  $\psi^{-1}$ ).  $\square$

From Proposition 3.9 and Lemma 3.14 we get a dual version of Proposition 3.7.

**Proposition 3.10.** *Let  $Q \in F_{C \cup X}$  be a non-orientable quadratic word. Then there exists an automorphism  $\psi \in \text{Aut}_{F_C}(F_{C \cup \text{Var}(Q)})$  such that  $Q^\psi$  is conjugate to a standard quadratic*

word equivalent to  $Q$  and for any  $x, y \in \text{Var}(Q)$  and  $c \in C$ , we have  $|x^{\psi^{-1}}|_y \leq 8 \text{genus}(Q)$  and  $|x^{\psi^{-1}}|_c \leq 4 \text{genus}(Q)$ .

Passing from quadratic words with formal coefficients to quadratic words  $Q \in G * F_X$  over a group  $G$  we can easily formulate a general result for an *arbitrary* coefficient group  $G$ .

**Corollary 3.11.** *Let  $Q \in G * F_X$  be quadratic word over an arbitrary group  $G$ . Then there are automorphisms  $\phi, \psi \in \text{Aut}_G(G * F_X)$  such that  $Q^\phi$  and  $Q^\psi$  are conjugate to standard quadratic words equivalent to  $Q$  and for any variable  $x$ ,*

$$|x^\phi|, |x^{\psi^{-1}}| \leq \begin{cases} 4n(Q) + 2c(Q) & \text{if } Q \text{ is orientable} \\ 8n^2(Q) + 4n(Q)c(Q) & \text{if } Q \text{ is non-orientable} \end{cases}$$

where  $c(Q)$  is the total length of coefficients of  $Q$  expressed in any left-invariant (e.g. word) metric on  $G$ .

*Proof.* For orientable  $Q$  this follows from Propositions 3.3 and 3.9. If  $Q$  is non-orientable then we have to use Propositions 3.7 and 3.10 and inequality  $\text{genus}(Q) \leq n(Q)$ .  $\square$

In the end of the section, we formulate several lemmas for later use. The first one was in fact proved in the proof of Proposition 3.9 (unlike the proposition, we do not make a passage to a conjugate element here).

**Lemma 3.12.** (i) *Let  $Q \in F_X$  be a coefficient-free orientable quadratic word. Then there is an automorphism  $\psi \in \text{Aut}(F_X)$  such that*

$$Q^\psi = [x_1, y_1] \dots [x_g, y_g]$$

*and  $|x_i^{\psi^{-1}}|_x, |y_i^{\psi^{-1}}|_x \leq 4$  for all  $i$  and  $x \in \text{Var}(Q)$ .*

(ii) *Let  $Q \in F_X$  be a coefficient-free non-orientable quadratic word. Then there is an automorphism  $\psi \in \text{Aut}(F_X)$  such that*

$$Q^\psi = x_1^2 \dots x_k^2 [x_{k+1}, y_{k+1}] \dots [x_{k+n}, y_{k+n}]$$

*and  $|x_i^{\psi^{-1}}|_x, |y_i^{\psi^{-1}}|_x \leq 4$  for all  $i$  and  $x \in \text{Var}(Q)$ .*

**Lemma 3.13.** *Let  $m \geq 1$  and  $\sigma$  be a permutation on  $\{1, \dots, m\}$ . Then there is an automorphism  $\psi$  of  $F_{\{c_1, \dots, c_m, z_1, \dots, z_m\}}$  such that  $c_i^\psi = c_i$  for all  $i$ , with the following properties:*

(i)

$$z_1^{-1} c_1 z_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m \xrightarrow{\psi} z_{\sigma(1)}^{-1} c_{\sigma(1)} z_{\sigma(1)} z_{\sigma(2)}^{-1} c_{\sigma(2)} z_{\sigma(2)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)} z_{\sigma(m)}.$$

(ii) *For any  $i$ , the image  $z_i^\psi$  of  $z_i$  has the form*

$$z_i^\psi = z_i z_{\sigma(j_1)}^{-1} c_{\sigma(j_1)} z_{\sigma(j_1)} \dots z_{\sigma(j_k)}^{-1} c_{\sigma(j_k)} z_{\sigma(j_k)}$$

*for some increasing sequence of indices  $1 \leq j_1 < \dots < j_k \leq m$ .*

*Proof.* Using the automorphism

$$z_i^{-1} c_i z_i z_j^{-1} c_j z_j \xrightarrow{(z_j \mapsto z_j z_i^{-1} c_i z_i)} z_j^{-1} c_j z_j z_i^{-1} c_i z_i$$

we can permute two neighboring factors of the form  $z_i^{-1} c_i z_i$ . Starting with  $z_{\sigma(m)}^{-1} c_{\sigma(m)} z_{\sigma(m)}$  we arrange factors  $z_i^{-1} c_i z_i$  to the right in the order as they should occur in the desired image

of  $z_1^{-1}c_1z_1 \dots z_m^{-1}c_mz_m$ . We get a triangular product of elementary automorphisms and all  $z_i^\psi$  have the required form as stated in (ii).  $\square$

**Lemma 3.14.** *Let*

$$Q_1 = x_1^2 \dots x_k^2 [x_{k+1}, x_{k+2}] \dots [x_{g-1}, x_g] z_1^{-1} c_1 z_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m$$

and

$$Q_2 = x_1^2 \dots x_r^2 [x_{r+1}, x_{r+2}] \dots [x_{g-1}, x_g] z_1^{-1} c_1^{\varepsilon_0} z_1 z_2^{-1} c_2^{\varepsilon_1} z_2 \dots z_m^{-1} c_m^{\varepsilon_m} z_m, \quad \varepsilon_1, \dots, \varepsilon_m = \pm 1,$$

be two non-orientable quadratic words of the same genus  $g$  with the same set of coefficient letters  $\{c_1, \dots, c_m\}$ . Then there is an automorphism  $\psi$  of  $F_{\{c_1, \dots, c_m\} \cup \text{Var}(Q_1)}$  such that  $c_i^\psi = c_i$  for all  $i$ , with the following properties:

- (i)  $Q_1^\psi$  is conjugate to  $Q_2$ .
- (ii) For any variable  $y \in \{x_1, \dots, x_g, z_1, \dots, z_m\}$ ,

$$|y^\psi|_Y \leq 4g, \quad |y^\psi|_Z \leq 4m \quad \text{and} \quad |y^\psi|_{c_i} \leq 2, \quad i = 1, \dots, m$$

where  $Y = \{x_1, \dots, x_g\}$  and  $Z = \{z_1, \dots, z_m\}$ .

*Proof.* We use the following set of automorphisms:

$$\begin{aligned} x^2[y, z] &\xrightarrow{\phi_1} y^2 z^2 x^2, & \phi_1 &= (x \mapsto y^2 z x y^{-1}, y \mapsto y z x y^{-1}, z \mapsto y x), \\ x^2[y, z] &\xrightarrow{\phi_2} [y, z] x^2, & \phi_2 &= (x \mapsto [y, z] x [y, z]^{-1}), \\ x^2 y^2 z^2 &\xrightarrow{\phi_3} [y, z] x^2, & \phi_3 &= (x \mapsto [y, z] x y, y \mapsto y^{-1} x^{-1} z^{-1}, z \mapsto z x), \\ x^2 z^{-1} c z &\xrightarrow{\phi_4} z^{-1} c^{-1} z x^2, & \phi_4 &= (x \mapsto z^{-1} c^{-1} z x, z \mapsto z x), \\ x^2 z^{-1} c z &\xrightarrow{\phi_5} z^{-1} c z x^2, & \phi_5 &= (z \mapsto z x^2). \end{aligned}$$

We pick up  $x_q^2$  with  $q = \min\{k, r\}$  and move it to the right of  $Q_1$  so that the square/commutator part becomes the same as in  $Q_2$  and each constant  $c_i$  gets the exponent sign  $\varepsilon_i$  as in  $Q_2$ . The bounds in (ii) are straightforward.  $\square$

#### 4. AN ELIMINATION PROCESS FOR QUADRATIC WORDS

There is a general approach to solutions of equations in free and similar groups which may be called *elimination process*. It usually deals with pairs  $(E, \alpha)$  where  $E = 1$  is an equation in a group  $F$  and  $\alpha$  is its solution. On the set of such pairs, a certain set of transformations is defined. The idea is to reduce step-by-step the cancellation which appears after substituting  $\alpha$  into  $E$  and then, starting from a given pair  $(E, \alpha)$ , to get a new pair  $(E_1, \alpha_1)$  of bounded complexity. In particular, solvability of equation  $E = 1$  is reduced to existence of a solution of bounded complexity of finitely many such equations  $E_1 = 1$  which can be algorithmically checked. To describe the solution set of  $E = 1$  one needs in addition to track transformations homomorphisms. (There are several ways to define them; for example, in the case of a free group  $F = F_A$  a transformation homomorphism from a pair  $(E, \alpha)$  to a pair  $(E_1, \alpha_1)$  can be viewed as an endomorphism  $\phi : F_{A \cup X} \rightarrow F_{A \cup X}$  such that  $E^\phi = E_1$  and  $\alpha = \phi \alpha_1$ .) For free groups, an elimination process has been applied in its full strength by Razborov [25] to provide a description of solutions of an arbitrary system of equations in free groups. In the case of quadratic equations, there is a much simpler version in [5, 6, 11].

In this section, we apply a version of the elimination process to pairs of the form  $(Q, \alpha)$  where  $Q \in F_X$  is a coefficient-free quadratic word and  $\alpha$  is an evaluation of variables of  $Q$ , i.e. a homomorphism  $F_X \rightarrow F_A$ . We use the process to give a bound on the size of *some* solution of a related quadratic equation and thus do not need to track transformation homomorphisms. A more sophisticated version (with tracking transformation homomorphisms) will be used in Section 6 to obtain bounds on the size of parametric solutions.

Let  $(Q, \alpha)$  be a pair where  $Q \in F_X$  is a coefficient-free quadratic word and  $\alpha : F_X \rightarrow F_A$  is an evaluation of variables of  $Q$ . We define several types of elementary transformations which carry  $(Q, \alpha)$  to a new pair  $(Q_1, \alpha_1)$ . In all cases, there will be a transformation homomorphism  $\phi \in \text{End}(F_X)$  such that  $Q_1 = Q^\phi$  (but we do not require that  $\alpha = \alpha_1\phi$ ). As one of the main inductive parameters, we take the length of a formal word  $Q[\alpha]$  obtained by substituting in  $Q$  values  $x^\alpha$  of all variables  $x \in \text{Var}(Q)$  represented by freely reduced words. It is straightforward to check that  $|Q_1[\alpha_1]| \leq |Q[\alpha]|$  for all elementary transformations introduced below.

*Degenerate transformation.* Application condition:  $x^\alpha = 1$  for some  $x \in \text{Var}(Q)$ . We do not change  $\alpha$  and apply the homomorphism  $\phi = (x \mapsto 1)$  to  $Q$ , so  $Q_1$  is obtained from  $Q$  by removing all occurrences of  $x$  and performing subsequent cancellation.

We introduce a natural notion related to this transformation. We say that a coefficient-free quadratic word  $Q_2$  is a *homomorphic image* of  $Q_1$  if there is an endomorphism  $\phi \in \text{End}(F_X)$  with  $Q_1^\phi = Q_2$ . The following fact is an easy exercise.

**Proposition 4.1.** *Let  $Q_1$  and  $Q_2$  be coefficient-free quadratic words. Then  $Q_2$  is a homomorphic image of  $Q_1$  if and only if either  $Q_1$  and  $Q_2$  are of the same orientability and  $\text{genus}(Q_2) \leq \text{genus}(Q_1)$  or  $Q_1$  is non-orientable,  $Q_2$  is orientable and  $\text{genus}(Q_2) < \frac{1}{2}\text{genus}(Q_1)$ .*

*Cancellation reduction.* Application condition: a non-trivial cancellation in  $Q[\alpha]$  occurs between the values of two neighboring variables  $x^\varepsilon$  and  $y^\delta$  ( $\varepsilon, \delta = \pm 1$ ). There are two cases.

*Case 1:*  $x \neq y$ . Then for some  $u, v$  and  $w \neq 1$ ,

$$(x^\varepsilon)^\alpha = uw \quad \text{and} \quad (y^\delta)^\alpha = w^{-1}v$$

where equality stands for graphical equality of words. We take a variable  $z \notin \text{Var}(Q)$  and define a transformation homomorphism  $\phi$  by

$$\phi = (x^\varepsilon \mapsto x^\varepsilon z, \quad y^\delta \mapsto z^{-1}y^\delta).$$

To define  $\alpha_1$  we set

$$(x^\varepsilon)^{\alpha_1} = u, \quad (y^\delta)^{\alpha_1} = v, \quad z^{\alpha_1} = w$$

and  $h^{\alpha_1} = h^\alpha$  for all other variables  $h \neq x, y, z$ .

*Case 2:*  $x = y$ . Then  $x^\varepsilon y^\delta$  becomes  $(x^\varepsilon)^2$  and for some  $u$  and  $w \neq 1$ ,

$$(x^\varepsilon)^\alpha = w^{-1}uw.$$

In a similar manner, we take

$$\phi = (x^\varepsilon \mapsto z^{-1}x^\varepsilon z)$$

and define  $\alpha_1$  by

$$(x^\varepsilon)^{\alpha_1} = u, \quad z^{\alpha_1} = w.$$

After application of this transformation we get a pair  $(Q_1, \alpha_1)$  which satisfies the strict inequality  $|Q_1[\alpha_1]| < |Q[\alpha]|$ .

*Splitting a variable.* Application condition:  $x$  is a variable in  $Q$  with  $x^\alpha \neq 1$ . We introduce a new variable  $y \notin \text{Var}(Q)$  and apply to  $Q$  the substitution  $\phi = (x \mapsto xy)$ . For  $\alpha_1$ , we take any homomorphism  $F_X \rightarrow F_A$  such that  $x^{\alpha_1}y^{\alpha_1} = x^\alpha$ , the product  $x^{\alpha_1}y^{\alpha_1}$  is reduced and  $\alpha_1$  coincides with  $\alpha$  on all variables other than  $x$  and  $y$ .

As an illustration, we prove a well known fact.

**Proposition 4.2** ([30]). *Let  $w \in F_A$  be a value of a coefficient-free quadratic word  $Q$ . Then there exists a homomorphic image  $R$  of  $Q$  and an evaluation  $\beta : F_X \rightarrow F_A$  such that  $x^\beta \neq 1$  for all  $x \in \text{Var}(R)$ , the word  $R[\beta]$  is freely reduced and equals to  $w$ .*

*Proof.* Let  $w = Q^\alpha$ . We start with the pair  $(Q, \alpha)$  and apply degenerate transformations and cancellation reductions until possible. For the resulting pair  $(R, \beta)$ ,  $R$  and  $\beta$  are as required.  $\square$

**Corollary 4.3.** (i) *If  $w \in F_A$  is a value of a coefficient-free orientable quadratic word of genus  $g$  then  $w$  is a product of  $g$  commutators  $[u_1, v_1] \dots [u_g, v_g]$  with  $|u_i|, |v_i| \leq 2|w|$  for all  $i$ .*  
(ii) *If  $w \in F_A$  is a value of a coefficient-free non-orientable quadratic word of genus  $g$  then  $w$  can be represented as a product  $u_1^2 \dots u_k^2 [u_{k+1}, u_{k+2}] \dots [u_{g-1}, u_g]$  where  $|u_i| \leq 2|w|$  for all  $i$ .*

*Proof.* Let  $Q$  be the quadratic word from the hypothesis and let  $w = v^{-1}w_1v$  where  $w_1$  is cyclically reduced. By Proposition 4.2 and Lemma 3.12, there is a standard orientable or semi-standard non-orientable quadratic word  $R$  and an evaluation  $\beta$  of variables in  $R$  such that  $R$  is a homomorphic image of  $Q$ ,  $R^\beta = w_1$  and  $|x^\beta| \leq 2|w_1|$  for every variable  $x \in \text{Var}(R)$ . If needed by we add to  $R$  extra commutators or squares with new variables  $y$  with values  $y^\beta = 1$  so that  $R$  becomes equivalent to  $Q$ . The required  $u_i$ 's and  $v_i$ 's are obtained by conjugating the values  $x^\beta$  with  $v$ .  $\square$

In what follows, we apply an elimination process to a pair  $(Q, \alpha)$  where  $\alpha$  is a solution of a standard quadratic equation written in the form

$$(1) \quad Q = z_1^{-1}c_1z_1z_2^{-1}c_2z_2 \dots z_m^{-1}c_mz_m.$$

Our goal is to find a solution of this equation satisfying the bound from Theorem 1.1. To do this, using elimination process we find first a short (in a certain sense) solution of an equivalent quadratic equation of the same form, for given fixed elements  $c_1, \dots, c_m \in F_A$ . Then, by using an “economical” automorphism from Section 3 we reduce the equation to the initial form thus producing the required “short” solution of the original equation (1).

Instead of keeping fixed the element  $Q^\alpha$  during transformations as in the proof of Proposition 4.2 we will keep the property that  $Q^\alpha$  is a product of conjugates of fixed elements  $c_1, \dots, c_m \in F_A$  (that is,  $w = u_1^{-1}c_1u_1 \dots u_m^{-1}c_mu_m$  for some  $u_1, \dots, u_m$ ), up to changing exponent signs of  $c_i$  in the non-orientable case.

**Definition 4.4.** Let  $w, c_1, c_2, \dots, c_m \in F_A$  be cyclically reduced elements of  $F_A$  and let  $w$  be a product of conjugates of  $c_1, \dots, c_m$ . We say that  $w$  is a *short* product of conjugates of  $c_1, \dots, c_m$  if  $w$ , viewed as a freely reduced cyclic word, does not have a form  $w = uhvh^{-1}$  where  $h$  is non-empty and the set  $\{c_i\}$  can be properly partitioned into two subsets  $\{c_{p_i}\}$  and  $\{c_{q_i}\}$  so that  $u$  is a product of conjugates of  $c_{p_i}$ 's and  $v$  is a product of conjugates of  $c_{q_i}$ 's.



**Definition 4.5.** (non-orientable version) Let  $w, c_1, c_2, \dots, c_m \in F_A$  be cyclically reduced elements of  $F_A$ . We say that  $w$  is a *unsigned product of conjugates* of  $c_1, \dots, c_m$  if  $w$  is a product of conjugates of  $c_1^{\varepsilon_1}, \dots, c_m^{\varepsilon_m}$  for some  $\varepsilon_1, \dots, \varepsilon_m = \pm 1$ .

We say that  $w$  is a *short unsigned product of conjugates* of  $c_1, \dots, c_m$  if  $w$ , viewed as a freely reduced cyclic word, does not have a form  $w = uhvh^{-1}$  where  $h$  is non-empty and the set  $\{c_i\}$  can be properly partitioned into two subsets  $\{c_{p_i}\}$  and  $\{c_{q_i}\}$  so that  $u$  is a unsigned product of conjugates of  $c_{p_i}$ 's and  $v$  is a unsigned product of conjugates of  $c_{q_i}$ 's.

It is not hard to prove that if  $w$  is a short or short unsigned product of conjugates of  $c_1, \dots, c_m$  then  $|w| \leq \sum |c_i|$ . This is essentially a consequence of the van Kampen lemma, see Lemma 5.5 below.

**Proposition 4.6.** Let  $c_1, c_2, \dots, c_m \in F_A$  and  $Q$  be a coefficient-free (orientable or non-orientable) quadratic word of genus  $g$ . Suppose that the quadratic equation

$$Q = z_1^{-1} c_1 z_1 \dots z_m^{-1} c_m z_m$$

has a solution in  $F_A$ .

- (i) If  $Q$  is orientable then a short product of conjugates of  $c_1, \dots, c_m$  is a product of at most  $g$  commutators in  $F_A$ .
- (ii) If  $Q$  is non-orientable then a short unsigned product of conjugates of  $c_1, \dots, c_m$  is a product of at most  $g$  squares in  $F_A$ .

*Proof.* We consider first the case of orientable  $Q$ .

Let  $\alpha$  be a solution of the equation from the hypothesis of the proposition. We describe a sequence of transformations starting with the pair  $(Q, \alpha)$ . Any moment we will have a coefficient-free quadratic word  $R$  and a homomorphism  $\beta : F_X \rightarrow F_A$  such that  $R$  is a homomorphic image of  $Q$  and  $R^\beta$  is a product of conjugates of  $c_1, \dots, c_m$ . For the inductive parameter that will be decreased during transformations, we take the pair  $(|R^\beta|, |R[\beta]|)$  with lexicographic ordering. Recall that  $|R^\beta|$  denotes the length of the freely reduced word representing  $R^\beta$  and  $|R[\beta]|$  denotes the length of the formal word  $R[\beta]$ .

If  $x^\beta = 1$  for some variable  $x \in \text{Var}(R)$  then we apply the degenerate transformation  $(x \mapsto 1)$  to  $(R, \beta)$  decreasing the number of variables in  $R$ . The element  $R^\beta$  and the formal word  $R[\beta]$  are not changed.

If  $R[\beta]$  has a cancellation then we apply a cancellation reduction so that  $R^\beta$  is not changed but the length of  $R[\beta]$  decreases.

If  $R[\beta]$  has a cyclic reduction then we change  $\beta$  by conjugating all the values  $x^\beta$  with the same element of  $F_A$  decreasing the length of  $R^\beta$ . Observe also that using this operation we can change  $R^\beta$  to its any cyclic shift not increasing parameter  $(|R^\beta|, |R[\beta]|)$ .

Thus we assume that  $R[\beta]$  is cyclically reduced (and hence is the cyclically reduced form of  $R^\beta$ ) and all variables  $x$  in  $R$  have non-trivial values  $x^\beta \neq 1$ .

Suppose that  $R^\beta$  is not a short product of conjugates of  $c_1, \dots, c_m$ , that is, up to a cyclic shift we have  $R^\beta = uhvh^{-1}$  where  $u, v$  and  $h \neq 1$  are as in Definition 4.4. Splitting variables if needed we may assume that  $h$  and  $h^{-1}$  are values of single variables in  $R$ , that is,

$$R = Ux_1^\varepsilon Vx_2^\delta, \quad U^\beta = u, \quad (x_1^\varepsilon)^\beta = h, \quad V^\beta = v, \quad (x_2^\delta)^\beta = h^{-1}.$$

We replace  $\beta$  by a new evaluation  $\beta_1$  such that the length of cyclically reduced form of  $R^\beta$  decreases and  $R^{\beta_1}$  is still a product of conjugates of  $c_1, \dots, c_m$ . After that we go back to the start of our procedure replacing  $R^{\beta_1}$  by its cyclically reduced form. There are two cases.

*Case 1:*  $x_1 = x_2$ . We define  $x^{\beta_1} = 1$  and leave the values of all other variables unchanged. Then  $R^{\beta_1} = (UV)^{\beta} = uv$  and by the condition of Definition 4.4,  $R^{\beta_1}$  is a product of conjugates of  $c_1, \dots, c_m$ .

*Case 2:*  $x_1 \neq x_2$ . Then  $x_1$  occurs either in  $U$  or in  $V$ . Suppose that  $U = U_1 x_1^{-\varepsilon} U_2$  (the case when  $x_1$  occurs in  $V$  is similar). We take  $x_1^{\beta_1} = ((U_2 U_1)^{\varepsilon})^{\beta}$  and  $y^{\beta_1} = y^{\beta}$  for all other  $y$ . Then

$$R^{\beta_1} = (U_2 U_1)^{\beta} v h^{-1} = U_2^{\beta} u (U_2^{\beta})^{-1} \cdot h v h^{-1}$$

and hence  $R^{\beta_1}$  is a product of conjugates of  $c_1, \dots, c_m$  but now we have  $|R^{\beta_1}| \leq |R^{\beta}| - 2|h|$ .

The description of the transformation sequence is finished. After finitely many steps we get a pair  $(R, \beta)$  such that  $R^{\beta}$  is a short product of conjugates of  $c_1, \dots, c_m$ . Since  $R$  is a homomorphic image of  $Q$ , it is a product of at most  $g$  commutators in  $F_X$ . This proves (i).

In the non-orientable case, the argument is similar with the difference that we keep  $R^{\beta}$  being an unsigned product of conjugates of  $c_1, \dots, c_m$ . (In Case 2, if  $x_1$  occurs in  $R$  twice with the same exponent  $\varepsilon$  then some exponent signs of  $c_1, \dots, c_m$  are changed after the transformation.)  $\square$

## 5. UNFOLDING LYNDON–VAN KAMPEN DIAGRAMS

Using Proposition 4.6 and automorphisms from Section 3 it is not difficult to get a bound on the size of the shortest solution of a standard quadratic equation in  $F_A$ . The bound is  $Nn(Q)c(Q)$  in the case of orientable  $Q$  and  $Nn(Q)^2 c(Q)$  in the case of non-orientable  $Q$ . To improve this bound by factor  $n(Q)$  we prove an extra statement. Informally speaking, it says that a Lyndon-van Kampen diagram can be unfolded in an economical way.

We recall some definitions and facts about Lyndon–van Kampen diagrams (or simply diagrams from now on for brevity).

By a diagram we mean a finite 2-dimensional cell complex  $\Delta$  embedded in the plane  $\mathbb{R}^2$  and endowed with a labelling function  $\lambda$  over an alphabet  $Y$ . The latter means that for any directed edge  $e$  of  $\Delta$  the label  $\lambda(e)$  is fixed which is either a letter in  $Y^{\pm 1}$  or is empty. For any two mutually inverse directed edges  $e$  and  $e^{-1}$ , we have  $\lambda(e^{-1}) = (\lambda(e))^{-1}$ . The labelling function is naturally extended to paths (viewed as sequences of directed edges) in the 1-skeleton  $\Delta^{(1)}$  of  $\Delta$ . The label  $\lambda(p)$  of a path  $p$  is a word in  $Y^{\pm 1}$  which we will often identify with an element of  $F_Y$ . We call the label of the boundary loop of a 2-cell  $D$  of  $\Delta$  the *boundary label* of  $\Delta$ , defined up to a cyclic shift.

We assume that all diagrams are connected and simply connected. We assume also that a diagram  $\Delta$  is endowed with a fixed *base vertex*  $\nu_0$  in the boundary of  $\Delta$  and, moreover, a boundary loop of  $\Delta$  at  $\nu_0$  is fixed. (In general, the boundary loop starting at a given boundary vertex may be not unique). The label of the fixed boundary loop of  $\Delta$  is called the *boundary label* of  $\Delta$ .

We admit that the boundary label of a 2-cell  $D$  of a diagram  $\Delta$  is the empty word or has the form  $yy^{-1}$ ,  $y \in Y$ . In this case we call  $D$  a *trivial* 2-cell of *vertex* or *edge* type, respectively. Otherwise a 2-cell is called *nontrivial*. The boundary label of a nontrivial 2-cell is always assumed to be cyclically reduced.

If words  $c_1, \dots, c_m$  are boundary labels of all nontrivial 2-cells of  $\Delta$  then we call  $\Delta$  a *diagram over the set*  $\{c_1, \dots, c_m\}$ . (To be more formal,  $\{c_1, \dots, c_m\}$  should be viewed as a multiset since we admit that some  $c_i$  are repeated.)

Let  $w, c_1, \dots, c_m \in F_Y$ . A variant of the van Kampen lemma says that  $w$  is a product of conjugates of  $c_1, \dots, c_m$  if and only if there is a diagram  $\Delta$  with labelling function over  $Y$  over the set  $\{c_1, \dots, c_m\}$ , with boundary label  $w$ . The proof can be found in [21, proof of Theorem V.1.1]. Note that cut off operations of spherical diagrams in the construction of  $\Delta$  can be avoided since we admit empty labels of edges and trivial 2-cells.

We need a precise description of the process of constructing a diagram from a representation of an element  $w$  as a product of conjugates of  $c_1, \dots, c_m$ . We start with describing several elementary operations applied to a given diagram  $\Delta$  which produce a new diagram  $\Delta_1$  over the same set  $\{c_1, \dots, c_m\}$ .

(T1) *Contracting a trivial edge.* If  $e$  is an edge with distinct endpoints and  $\lambda(e) = 1$  then we contract  $e$  into a vertex.

(T2) *Contracting a trivial 2-cell.* Let  $D$  be a trivial 2-cell of  $\Delta$ . Assume that either  $D$  has vertex type and the boundary loop consists of one edge or  $D$  has edge type and the boundary loop of  $D$  consists of exactly 2 edges and is labelled  $yy^{-1}$ ,  $y \in Y$ . Then we contract  $D$  to a vertex or to an edge labelled  $y$ , respectively.

We admit also inverse operations  $(T1)^{-1}$  and  $(T2)^{-1}$ , *introducing a trivial edge* and *introducing a trivial 2-cell*, respectively. We call operations (T1), (T2) and their inverses *trivial transformations*. We introduce another two types of elementary operations which we call *elementary reductions*.

(R1) *Folding boundary edges.* Assume that two distinct boundary directed edges  $e_1$  and  $e_2$  of  $\Delta$  have the same label, a common initial vertex  $\nu$  and distinct terminal vertices. We assume furthermore that the path  $(e_1 e_2)^{\pm 1}$  occurs in the boundary loop of  $\Delta$ . (This is not always true in the case when  $\nu$  is the base vertex of  $\Delta$ .) Then we perform folding of  $e_1$  and  $e_2$  into a single edge.

(R2) *Removal of a leaf edge.* Let  $e$  be an edge of  $\Delta$  with an endpoint  $\nu$  of valence 1. We assume that  $\nu$  is not the base vertex of  $\Delta$ . Then we remove  $e$  and its endpoint  $\nu$  from  $\Delta$ . Note that  $e$  does not belong to the boundary of a nontrivial 2-cell of  $\Delta$  since boundary labels of nontrivial 2-cells are assumed to be cyclically reduced. Therefore, this operation either reduces cancellation in the boundary label of  $\Delta$  or changes a trivial 2-cell of edge type to a trivial 2-cell of vertex type.

By definition, for all the operations introduced, the base vertex and the boundary loop of  $\Delta_1$  are inherited in the natural way from those of  $\Delta$ .

Observe that (T1) and (T2) do not change the boundary label  $w$  of  $\Delta$ , (R1) reduces a cancellation in  $w$  and (R2) either does not change  $w$  or reduces a cancellation in  $w$ .

It is easy to see that any cancellation in the boundary label of  $\Delta$  can be reduced by either (R1) or (R2) after a sequence of trivial transformations  $(T1)^{\pm 1}$  and  $(T2)^{\pm 1}$ .

**Definition 5.1.** We say that a diagram  $\Delta_2$  is obtained by *folding* from a diagram  $\Delta_1$  if it is the result of application of a sequence of trivial transformations and elementary reductions so that the boundary label of  $\Delta_2$  is freely reduced. We say also that  $\Delta_1$  is obtained by *unfolding* from  $\Delta_2$ .

For technical convenience, we assume that after folding, diagram  $\Delta_2$  always undergoes the following *tightening* procedure: First, contract any trivial 2-cell whenever it can be contracted by using (T2) and any sequence of operations  $(T1)^{\pm 1}$ . Second, contract all trivial edges whenever possible.

Below we use the following properties of *tight* diagrams as in Definition 5.1. The proof of the following lemma is an easy exercise and left to the reader.

**Lemma 5.2.** *Let  $\Delta$  be a tight diagram. Then the following assertions are true.*

- (i) *All trivial 2-cells of  $\Delta$  have edge type. If a nontrivial edge  $e$  occurs in the boundary loop of such a 2-cell  $D$  then both  $e$  and  $e^{-1}$  occur in the boundary loop of  $D$  (so the union of  $D$  and  $e$  is an annulus). In particular, no non-trivial boundary edge of  $\Delta$  belongs to the boundary of a trivial 2-cell.*
- (ii) *Any simple loop in the 1-skeleton  $\Delta^{(1)}$  of  $\Delta$  bounds a subdiagram which has at least one nontrivial 2-cell.*

With a formal product

$$v = u_1^{-1}c_1u_1 \dots u_m^{-1}c_mu_m$$

we associate a *rose diagram*  $\Delta_0$  over the set  $\{c_1, \dots, c_m\}$  labelled with (perhaps, not reduced) word  $v$ , see Fig. 8. Using folding we can produce a new diagram  $\Delta$  whose boundary label is the freely reduced form of  $v$ . (Existence of such a diagram  $\Delta$  is essentially the main part of the van Kampen lemma.)

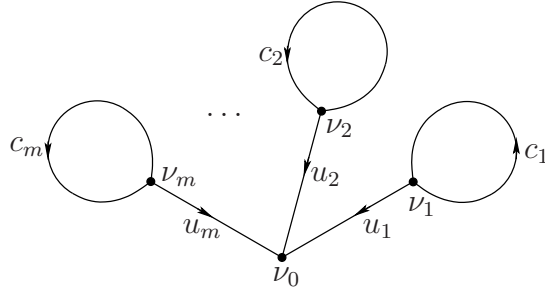


FIGURE 8.

We can consider also an inverse process: given a diagram  $\Delta$  over a set  $\{c_1, \dots, c_m\}$  with boundary label  $w$  we can unfold  $\Delta$  to get a rose diagram  $\Delta_0$  and its associated factorization

$$w = u_1^{-1}c'_{\sigma(1)}u_1 \dots u_m^{-1}c'_{\sigma(m)}u_m$$

of  $w$  in the free group  $F_Y$  where  $\sigma$  is a permutation on the set  $\{1, \dots, m\}$  and  $c'_i$  is a cyclic shift of  $c_i$ . The following proposition describes the relationship between two factorizations obtained in this way.

**Proposition 5.3.** *Let  $\Delta$  be a diagram with labeling function over  $Y$ , with boundary label  $w$ . Let  $\Delta_1$  and  $\Delta_2$  be two rose diagrams obtained from  $\Delta$  by unfolding and let*

$$w = u_1^{-1}c_1u_1 \dots u_m^{-1}c_mu_m = v_{\sigma(1)}^{-1}c'_{\sigma(1)}v_{\sigma(1)} \dots v_{\sigma(m)}^{-1}c'_{\sigma(m)}v_{\sigma(m)}$$

*be the associated factorizations of  $w$  in  $F_Y$  where  $\sigma$  is a permutation on the set  $\{1, \dots, m\}$  and  $c'_i$  is a cyclic shift of  $c_i$ . Then there is an  $F_Y$ -automorphism  $\phi$  of  $F_{Y \cup \{z_1, \dots, z_m\}}$  such that*

$$(z_1^{-1}c_1z_1 \dots z_m^{-1}c_mz_m)^\phi = z_{\sigma(1)}^{-1}c'_{\sigma(1)}z_{\sigma(1)} \dots z_{\sigma(m)}^{-1}c'_{\sigma(m)}z_{\sigma(m)}$$

and the following diagram is commutative

$$\begin{array}{ccc} F_{Y \cup \{z_1, \dots, z_m\}} & \xrightarrow{\phi} & F_{Y \cup \{z_1, \dots, z_m\}} \\ & \searrow \alpha \quad \swarrow \beta & \\ & F_Y & \end{array}$$

where  $\alpha$  and  $\beta$  are the corresponding evaluations of variables  $z_i$ :

$$z_i^\alpha = u_i, \quad z_i^\beta = v_i \quad (i = 1, \dots, m).$$

*Proof.* Let  $\nu_0$  be the base vertex of  $\Delta_1$  and  $\nu_1, \dots, \nu_m$  be vertices where disks with boundary labels  $c_i$  are attached to the other part of  $\Delta_1$  (shown in Fig. 8). Let  $z_i$  be the simple path between  $\nu_i$  and  $\nu_0$  and  $d_i$  the boundary loop of the corresponding 2-cell at  $\nu_i$  (which is labelled  $c_i$ ). We keep similar notations  $\nu'_i, z'_i$  and  $d'_i$  for  $\Delta_2$ .

Unfolding transformations  $\Delta \rightarrow \Delta_i$  induce a homotopy equivalence  $\psi$  between 1-skeletons  $\Delta_1^{(1)}$  and  $\Delta_2^{(1)}$  such that  $\psi(\nu_0) = \psi(\nu'_0)$ . In particular, we have the induced isomorphism  $\pi_1(\Delta_1^{(1)}, \nu_0) \rightarrow \pi_1(\Delta_2^{(1)}, \nu'_0)$  of fundamental groups. Moreover, from the construction of elementary operations it is easy to see that  $\psi$  satisfies the following conditions:

- (i) For any loop  $p$  at  $\nu_0$ , the image  $\psi(p)$  is a loop at  $\nu'_0$  labelled with a word representing the same element of  $F_Y$ .
- (ii)  $\psi(z_i^{-1}d_i z_i)$  has the form  $w_i^{-1}d'_i w_i$  for some  $w_i$  up to homotopy in  $\Delta_2^{(1)}$  rel  $\nu'_0$ .
- (iii) The image  $\psi(\ell)$  of the boundary loop  $\ell$  of  $\Delta_1$  is homotopic in  $\Delta_2^{(1)}$  rel  $\nu'_0$  to the boundary loop of  $\Delta_2$ .

The path  $z'_i{}^{-1}w_i$  is a loop at  $\nu'_0$ , so it has an expression (up to homotopy rel  $\nu'_0$ )

$$z'_i{}^{-1}w_i = f_i(z'_1{}^{-1}d'_1 z'_1, \dots, z'_m{}^{-1}d'_m z'_m)$$

in terms of the generators  $z'_i{}^{-1}d'_i z'_i$  of  $\pi_1(\Delta_2, \nu'_0)$ . We define formally a homomorphism  $\hat{\psi} : F_{\{d_1, \dots, d_m, z_1, \dots, z_m\}} \rightarrow F_{\{d'_1, \dots, d'_m, z'_1, \dots, z'_m\}}$  by

$$\hat{\psi}(d_i) = d'_i \quad \text{and} \quad \hat{\psi}(z_i) = z'_i f_i.$$

The definition implies that for any loop  $p$  at  $\nu_0$ , we have  $\psi(p) = \hat{\psi}(p)$  up to homotopy rel  $\nu'_0$ . From this we conclude that all  $z'_i{}^{-1}d'_i z'_i$  belong to the image of  $\hat{\psi}$  and hence  $\hat{\psi}$  is in fact an isomorphism. From (iii) we get

$$z_1^{-1}d_1 z_1 \dots z_m^{-1}d_m z_m \xrightarrow{\hat{\psi}} z'_{\sigma(1)}{}^{-1}d'_{\sigma(1)} z'_{\sigma(1)} \dots z'_{\sigma(m)}{}^{-1}d'_{\sigma(m)} z'_{\sigma(m)}$$

Now we adjust  $\hat{\psi}$  to get the required  $\phi$ . By (i) and (ii), we have

$$\lambda(z_i) = g_i \lambda(w_i) \quad \text{for some } g_i \in F_Y \text{ with } g_i^{-1}c_i g_i = c'_i$$

We define  $\phi$  by

$$\phi(z_i) = g_i z_i f_i(z_1^{-1}c'_1 z_1, \dots, z_m^{-1}c'_m z_m)$$

It is easy to see that  $\phi$  satisfies all the required conditions. □

To formulate the main result of the section, we need two more definitions. Let  $e$  be a nontrivial edge of a diagram  $\Delta$ . We call  $e$  a *tree edge* if both  $e$  and  $e^{-1}$  occur in the boundary loop of  $\Delta$  (or, in other words,  $\Delta$  splits into two disjoint subdiagrams after removal of  $e$ ). We call  $e$  *tubular* if  $e$  and  $e^{-1}$  occur in the boundary loop of a trivial 2-cell of  $\Delta$ .

**Proposition 5.4.** *Let  $w, c_1, c_2, \dots, c_m$  be cyclically reduced elements of a free group  $F_Y$ . Let  $\Delta$  be a diagram with boundary label  $w$  over the set  $\{c_1, \dots, c_m\}$  and let  $N$  denote the total length of labels of tree and tubular edges of  $\Delta$ . Assume that  $\Delta$  is tight as in Definition 5.1.*

*Then there is rose diagram obtained from  $\Delta$  by unfolding such that for the associated factorization*

$$w = v_1^{-1} c'_{\sigma(1)} v_1 \dots v_m^{-1} c'_{\sigma(m)} v_m$$

*where  $\sigma$  is a permutation on the indices  $\{1, \dots, m\}$  and  $c'_i$  is a cyclic shift of  $c_i$ , the following is true:*

- (i)  $|v_i| \leq \sum_{i=1}^m |c_i| + 2N$  for all  $i$ .
- (ii) for any increasing sequence  $1 \leq t_1 < t_2 < \dots < t_k \leq m$  of indices  $t_i$ ,

$$|v_{t_1}^{-1} c'_{\sigma(t_1)} v_{t_1} \dots v_{t_k}^{-1} c'_{\sigma(t_k)} v_{t_k}| \leq \sum_{i=1}^m |c_i| + 2N.$$

*Proof.* Suppose that  $e$  is an edge in the interior of  $\Delta$  such that at least one of the endpoints of  $e$  belongs to the boundary of  $\Delta$ . Then we can apply to  $\Delta$  an unfolding operation  $(R1)^{-1}$  replacing  $e$  by two new boundary edges.

Starting from  $\Delta$ , we perform recursively unfoldings of all interior edges. Then we contract all trivial 2-cells using (T1) and (T2). The resulting diagram  $\tilde{\Delta}$  has no interior edges and no trivial 2-cells. For convenience, we further introduce new trivial edges where needed, so that each vertex of  $\tilde{\Delta}$  gets valence at most 3 (see Fig. 9)

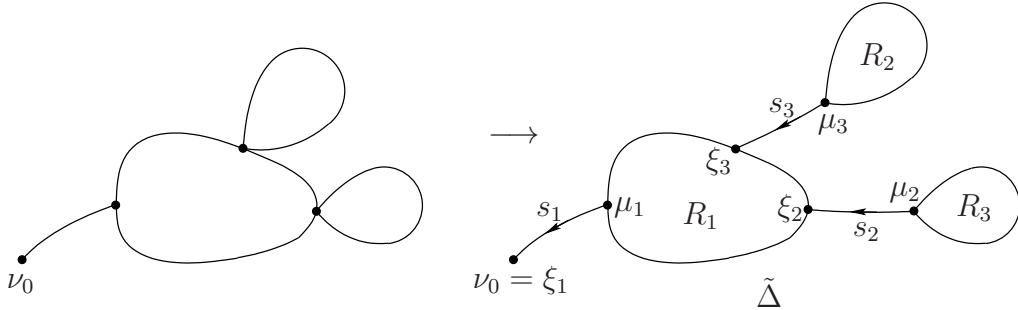


FIGURE 9.

We claim that  $\tilde{\Delta}$  has exactly  $N$  tree edges. Indeed, since  $\Delta$  is tight, by Lemma 5.2(i) any nontrivial interior edge of  $\Delta$  either occurs in the boundary loop of a nontrivial 2-cell or is tubular. Then every tree edge of  $\tilde{\Delta}$  comes either from a tree edge of  $\Delta$  (which is not changed in the transformation) or from a tubular edge  $e$  of  $\Delta$  (which becomes a tree edge after unfolding  $e$  and contracting the trivial 2-cell  $D$  with  $\partial D \supset e$ ).

Let  $\nu_1$  be the base vertex of  $\tilde{\Delta}$  and  $\ell$  be the boundary loop of  $\tilde{\Delta}$ . We enumerate cells  $R_1, \dots, R_m$  of  $\tilde{\Delta}$  in the order as their boundary edges meet first in  $\ell$ . Let

$$\{v_0\} = \Delta_0 \subset \Delta_1 \subset \dots \subset \Delta_m = \tilde{\Delta}$$

be a sequence of subdiagrams of  $\tilde{\Delta}$  where each  $\Delta_i$  is obtained from  $\Delta_{i-1}$  by attaching the topological closure of  $R_i$  and a segment  $s_i$  joining  $R_i$  with  $\Delta_{i-1}$ . We view  $s_i$  as a path from a vertex  $\mu_i$  in  $\partial R_i$  to a vertex  $\xi_i$  in  $\partial \Delta_{i-1}$ . The boundary loop of  $R_i$  at  $\mu_i$  is labelled with



a cyclic shift  $c'_{\sigma(i)}$  of  $c_{\sigma(i)}$  where  $\sigma$  is a permutation on  $\{1, \dots, m\}$ . Let  $r_i$  be the terminal segment of the boundary loop of  $\Delta_{i-1}$  starting at  $\xi_i$ . Denote  $v_i = \lambda(s_i r_i)$ . Then we have

$$w = v_1^{-1} c'_{\sigma(1)} v_1 \dots v_m^{-1} c'_{\sigma(m)} v_m$$

and from the construction we can easily see that the rose diagram associated with this factorization is obtained by unfolding from  $\tilde{\Delta}$  successively along paths  $r_m, r_{m-1}, \dots, r_1$  each time slicing off one 2-cell  $D_i$ .

Observe that  $s_i r_i$  passes through every tree edge of  $\tilde{\Delta}$  at most twice and through every other edge of  $\tilde{\Delta}$  at most once. This implies

$$|u_i| \leq \sum_{i=1}^m |c_i| + 2N.$$

Let us prove (ii). Let  $1 \leq t_1 < t_2 < \dots < t_k \leq m$  be an increasing sequence of indices  $t_i$ . Let  $h_i$  be the component of the intersection  $\partial R_i \cap \ell$  that meets first in  $\ell$ . We consider a diagram  $\Delta'$  obtained by removal from  $\tilde{\Delta}$  all 2-cells  $R_i$  and all arcs  $h_i$  with  $i \neq t_j$ . It is easy to see that the boundary label of  $\Delta'$  is equal to  $\prod_{i=1}^k v_{t_i}^{-1} c'_{\sigma(t_i)} v_{t_i}$ . Hence

$$\left| \prod_{i=1}^k v_{t_i}^{-1} c'_{\sigma(t_i)} v_{t_i} \right| \leq \sum_{i=1}^m |c_i| + 2N$$

□

For the proof of Theorem 1.1, we will use Proposition 5.4 in the special case when  $w$  is a short product of conjugates of elements  $c_1, \dots, c_m$  in the sense of Definition 4.4. We start with the following simple observation.

**Lemma 5.5.** *Let  $w, c_1, c_2, \dots, c_m$  be cyclically reduced elements of a free group  $F_Y$  and  $w$  be a short product of conjugates of  $c_1, \dots, c_m$ . Then any diagram  $\Delta$  over  $\{c_1, \dots, c_m\}$  with boundary label  $w$  has no tree edges. In particular,  $|w| \leq \sum_i |c_i|$ .*

*Proof.* If  $e$  is a tree edge of  $\Delta$  then the boundary loop of  $\Delta$ , up to a cyclic shift, has the form  $peqe^{-1}$  where  $p$  and  $q$  are loops which bound two subdiagrams  $\Delta_1$  and  $\Delta_2$  of  $\Delta$ . Since  $w$  is cyclically reduced, the labels of  $p$  and  $q$  are nontrivial elements of  $F_Y$  and hence both  $\Delta_1$  and  $\Delta_2$  have nontrivial 2-cells. We get a factorization of  $w$  as in Definition 4.4 which contradicts the assumption that  $w$  is short.

To prove the second statement, we take a tight diagram  $\Delta$  over  $\{c_1, \dots, c_m\}$  with boundary label  $w$ . By the first statement and Lemma 5.2(i), any boundary edge of  $\Delta$  belongs to the boundary of a nontrivial 2-cell. The total length of their boundary labels is  $\sum_i |c_i|$ . □

**Corollary 5.6.** *Let  $w, c_1, c_2, \dots, c_m$  be cyclically reduced elements of a free group  $F_Y$  and let  $w$  be a short product of conjugates of  $c_i$ 's. Then, up to re-enumeration and cyclic shifts of  $c_i$ 's, there exists a factorization*

$$w = u_1^{-1} c_1 u_1 \dots u_m^{-1} c_m u_m.$$

such that  $|u_i| \leq \sum_{i=1}^m |c_i|$  and for any increasing sequence of indices  $1 \leq t_1 < t_2 < \dots < t_k \leq m$ ,

$$|u_{t_1}^{-1} c_{t_1} u_{t_1} \dots u_{t_k}^{-1} c_{t_k} u_{t_k}| \leq \sum_{i=1}^m |c_i|.$$

*Proof.* Take any diagram  $\Delta$  over  $\{c_1, \dots, c_m\}$  with boundary label  $w$ . Assume that  $e$  is a tubular edge of  $\Delta$ . Let  $D$  be the trivial 2-cell of  $\Delta$  with boundary loop  $peqe^{-1}$  where  $\lambda(p) = \lambda(q) = 1$ . One of the loops  $p$  or  $q$  bounds a subdiagram  $\Delta_1$  of  $\Delta$ . Then we can remove the annulus  $D \cup e$  from  $\Delta$  replacing the subdiagram  $\Delta_1 \cup D \cup e$  by  $\Delta_1$ .

Continuing this process we may assume that  $\Delta$  has no tubular edges. It remains to apply Lemma 5.5 and Proposition 5.4.  $\square$

*Proof of Theorem 1.1.* Assume that a quadratic equation  $Q = 1$  is solvable in  $F_A$ . Let  $\psi \in \text{Aut}_{F_A}(F_{AUX})$  be an automorphism given by Proposition 3.3 or Proposition 3.5 such that  $Q^\psi$  is conjugate to a standard orientable or semi-standard non-orientable quadratic word  $R$ . If  $\alpha$  is a solution of  $R = 1$  then  $\psi\alpha$  is a solution of  $Q = 1$  and for any  $x \in \text{Var}(Q)$ ,

$$|x^{\psi\alpha}| \leq 4n(Q) \max_{x \in \text{Var}(Q)} |x^\alpha| + 2c(Q).$$

We have  $n(R) \leq n(Q)$  and  $c(R) \leq c(Q)$  and we can always assume that  $x^\alpha = 1$  for every  $x \in \text{Var}(Q) \setminus \text{Var}(R)$ . This implies that the statement of the theorem for  $Q$  follows from the statement for  $R$ . Therefore, it is sufficient to prove the theorem in the case of a standard or semi-standard  $Q$ . Let

$$Q = [x_1, y_1][x_2, y_2] \dots [x_g, y_g] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m$$

or

$$Q = x_1^2 x_2^2 \dots x_k^2 [x_{k+1}, x_{k+2}] \dots [x_{g-1}, x_g] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m.$$

We can further assume that all  $c_i$  are cyclically reduced. Indeed, in case of general  $c_i$ 's we consider the equation  $\bar{Q} = 1$  where each  $c_i$  is replaced with its cyclically reduced form  $\bar{c}_i = u_i^{-1} c_i u_i$ . A solution of  $Q = 1$  can be obtained from a solution of  $\bar{Q} = 1$  by the substitution

$$(x_i \mapsto u_1 x_i u_1^{-1}, y_i \mapsto u_1 y_i u_1^{-1}, z_i \mapsto u_i z_i u_i^{-1} \text{ for all } i).$$

It is easy to see that the statement of the theorem for  $Q = 1$  follows from the statement for  $\bar{Q} = 1$ .

Now consider two cases.

*Case 1:*  $Q$  is standard orientable, i.e.

$$Q = [x_1, y_1][x_2, y_2] \dots [x_g, y_g] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m.$$

By Proposition 4.6, there exists a solution  $\beta$  of the equation

$$[x_1, y_1][x_2, y_2] \dots [x_g, y_g] z_1^{-1} c_1 z_1 \dots z_m^{-1} c_m z_m = 1$$

such that the cyclically reduced form  $w$  of  $(z_1^{-1} c_1 z_1 \dots z_m^{-1} c_m z_m)^\beta$  is a short product of conjugates of  $c_i$ 's. In particular,  $|w| \leq c(Q)$  by Lemma 5.5. By Corollary 5.6, there exist a permutation  $\sigma$  on  $\{1, \dots, m\}$  and elements  $u_1, \dots, u_m \in F_A$  of length  $|u_i| \leq 2c(Q)$  such that

$$w = u_1^{-1} c_{\sigma(1)} u_1 \dots u_m^{-1} c_{\sigma(m)} u_m$$

and for any increasing sequence of indices  $1 \leq t_1 < t_2 < \dots < t_k \leq m$ ,

$$|u_{t_1}^{-1} c_{\sigma(t_1)} u_{t_1} \dots u_{t_k}^{-1} c_{\sigma(t_k)} u_{t_k}| \leq c(Q).$$

This implies by Lemma 3.13 that there are elements  $v_1, \dots, v_m$  of length  $|v_i| \leq 3c(Q)$  such that

$$w = v_1^{-1} c_1 v_1 \dots v_m^{-1} c_m v_m.$$

By Corollary 4.3,  $w$  can be represented as a product of  $g$  commutators of elements of length at most  $2c(Q)$ . Hence we get a solution  $\gamma$  of the equation

$$(2) \quad [x_1, y_1][x_2, y_2] \dots [x_g, y_g] z_1^{-1} c_1 z_1 \dots z_m^{-1} c_m z_m = 1$$

such that  $|x_i^\gamma|, |y_i^\gamma| \leq 2c(Q)$  and  $|z_i^\gamma| \leq 3c(Q)$ . To get a solution of the original equation  $Q = 1$ , we eliminate  $z_1$  by applying to the left-hand side of (2) the automorphism  $\psi \in \text{Aut}(F_{A \cup X})$  defined by

$$\psi = (x_i \mapsto z_1^{-1} x_i z_1, y_i \mapsto z_1^{-1} y_i z_1 \text{ for all } i, z_j \mapsto z_j z_1 \text{ for } j > 1).$$

This gives a solution  $\psi^{-1}\gamma$  of  $Q = 1$  with  $|x^{\psi^{-1}\gamma}| \leq 8c(Q)$  for all  $x$ .

*Case 2:*  $Q$  is semi-standard non-orientable, i.e.

$$Q = x_1^2 x_2^2 \dots x_k^2 [x_{k+1}, x_{k+2}] \dots [x_{g-1}, x_g] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m.$$

Similarly to Case 1, for some  $\varepsilon_1, \dots, \varepsilon_m \in \{-1, 1\}$  we find a solution  $\gamma$  of another semi-standard equation

$$x_1^2 \dots x_r^2 [x_{r+1}, x_{r+2}] \dots [x_{g-1}, x_g] z_1^{-1} c_1^{\varepsilon_1} z_1 \dots z_{m-1}^{-1} c_m^{\varepsilon_m} z_m = 1$$

where  $|x_i^\gamma| \leq 2c(Q)$  and  $|z_i^\gamma| \leq 3c(Q)$  for all  $i$ . By Lemma 3.14, there is an automorphism  $\phi \in \text{Aut}(F_{A \cup X})$  carrying its left-hand side, up to conjugation, to a word

$$Q_1 = x_1^2 \dots x_k^2 [x_{k+1}, x_{k+2}] \dots [x_{g-1}, x_g] z_1^{-1} c_1 z_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m$$

which produces a solution  $\gamma_1 = \phi^{-1}\gamma$  of  $Q_1 = 1$  with

$$|x^{\gamma_1}| \leq (8g + 12m + 2)c(Q) \quad \text{for all } x.$$

Finally, we eliminate  $z_1$  as in the orientable case which at most triples the bound, and observe that  $8g + 12m + 2 \leq 12n(Q)$ .  $\square$

## 6. BOUNDING PARAMETRIC SOLUTIONS

In this section we prove Theorem 1.2.

As defined in Section 1, by a parametric solution of an equation  $E = 1$  in a free group  $F_A$  we mean an  $F_A$ -homomorphism  $\eta : F_{A \cup \text{Var}(E)} \rightarrow F_{A \cup T}$  such that  $E^\eta = 1$  where  $T$  is a set of parameters. Here instead of the “big” group  $F_{A \cup X}$  we consider the group  $F_{A \cup \text{Var}(E)}$  involving only variables occurring in  $Q$ . It will be convenient to change this point of view by introducing formal sets of variables for equations. We assume that an equation  $E = 1$  is endowed with a formal finite set of variables  $V \subset X$  such that  $V \supseteq \text{Var}(E)$  (in other words, we admit now fictitious variables  $x \in V$  not occurring in  $E$ ). A parametric solution of such an equation  $(E = 1, V)$  is then an  $F_A$ -homomorphism  $\beta : F_{A \cup V} \rightarrow F_{A \cup T}$  such that  $E^\beta = 1$ . If  $V = \text{Var}(E)$  then we get equations and their parametric solutions in the initial sense.

Transformations of equations are no longer  $F_A$ -automorphisms (or  $F_A$ -endomorphisms if degenerate transformations are allowed) of the big free group  $F_{A \cup X}$  but homomorphisms  $\phi : F_{A \cup V} \rightarrow F_{A \cup V_1}$  where  $V, V_1 \subset X$  are finite sets of variables. Since we want  $\phi$  to be “potentially invertible” we require that  $\phi$  be a monomorphism. Moreover, we will require that the condition given by the following definition should be satisfied.

**Definition 6.1.** We call an  $F_A$ -monomorphism  $\phi : F_{A \cup V} \rightarrow F_{A \cup V_1}$  *primitive* if the image of  $\phi$  is a free factor of  $F_{A \cup V_1}$ .

The main ingredient to the proof of Theorem 1.2 is the following proposition which may be viewed as an advanced form of Proposition 4.6.

**Proposition 6.2.** *Let  $\{c_1, \dots, c_m\}$  be a finite set of cyclically reduced elements of  $F_A$ . Suppose that  $\eta : F_{A \cup \text{Var}(Q) \cup Z} \rightarrow F_{A \cup T}$  is a parametric solution of a quadratic equation in  $F_A$  of the form*

$$Q = z_1^{-1} c_1 z_1 \dots z_m^{-1} c_m z_m$$

where  $Q$  is a coefficient-free quadratic word and  $Z = \{z_1, \dots, z_m\}$ .

Then there is a coefficient-free quadratic word  $R$  equivalent to  $Q$ , a finite set of variables  $V \supseteq \text{Var}(R)$  and a parametric solution  $\theta : F_{A \cup V \cup Z} \rightarrow F_{A \cup T}$  of an equation

$$R = z_{\sigma(1)}^{-1} c_{\sigma(1)}^{\varepsilon_1} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)}^{\varepsilon_m} z_{\sigma(m)}$$

where all  $\varepsilon_i = 1$  if  $Q$  is orientable,  $\varepsilon_i \in \{-1, +1\}$  if  $Q$  is non-orientable and  $\sigma$  is a permutation on  $\{1, 2, \dots, m\}$ , such that the following assertions are true:

- (i) *There is a primitive  $F_A$ -monomorphism  $\phi : F_{A \cup \text{Var}(Q) \cup Z} \rightarrow F_{A \cup V \cup Z}$  and an  $F_A$ -endomorphism  $\omega \in \text{End}_{F_A}(F_{A \cup T})$  such that*

$$(Q^{-1} z_1^{-1} c_1 z_1 \dots z_m^{-1} c_m z_m)^\phi$$

*is conjugate to*

$$R^{-1} z_{\sigma(1)}^{-1} c_{\sigma(1)}^{\varepsilon_1} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)}^{\varepsilon_m} z_{\sigma(m)}$$

*and the following diagram is commutative:*

$$\begin{array}{ccc} F_{A \cup \text{Var}(Q) \cup Z} & \xrightarrow{\phi} & F_{A \cup V \cup Z} \\ \eta \downarrow & & \downarrow \theta \\ F_{A \cup T} & \xleftarrow{\omega} & F_{A \cup T} \end{array}$$

- (ii) *Let  $\bar{R}$  denote the word obtained by removing from  $R$  all variables  $x$  with  $x^\theta = 1$  and performing all subsequent cancellations. Then  $\bar{R}[\theta]$  is cyclically reduced (and hence  $\bar{R}[\theta]$  is the cyclically reduced form of  $R^\theta$ ; recall that  $W[\theta]$  denotes the formal word obtained after substitution in  $W$  of values  $x^\theta$  of all variables  $x$ ).*
- (iii) *There is a Lyndon–van Kampen diagram  $\Delta$  with boundary label  $\bar{R}^\theta$  over the set  $\{c_1^{\varepsilon_1}, \dots, c_m^{\varepsilon_m}\}$  folded from the rose diagram associated with factorization*

$$(3) \quad \bar{R}^\theta = (z_{\sigma(1)}^\theta)^{-1} c_{\sigma(1)}^{\varepsilon_1} z_{\sigma(1)}^\theta \dots (z_{\sigma(m)}^\theta)^{-1} c_{\sigma(m)}^{\varepsilon_m} z_{\sigma(m)}^\theta$$

*such that any tree or tubular edge of  $\Delta$  is labelled by a single parameter letter and the total number of tree and tubular edges is less than  $m$ .*

*Proof.* We consider first the case of orientable  $Q$  (so all  $\varepsilon_i$  are 1).

We describe a certain transformation process. At any moment, we will have the following data:

- A quadratic word  $R^*$  of the form

$$R^* = R^{-1} z_{\sigma(1)}^{-1} c_{\sigma(1)} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)} z_{\sigma(m)}$$

where  $R$  is a coefficient-free quadratic word equivalent to  $Q$ ;

- A finite set  $V$  of variables such that  $V \supseteq \text{Var}(R)$ .
- A parametric solution  $\theta : F_{A \cup V \cup Z} \rightarrow F_{A \cup T}$  of the equation  $R^* = 1$ .

- A primitive  $F_A$ -monomorphism  $\phi : F_{A \cup \text{Var}(Q) \cup Z} \rightarrow F_{A \cup V \cup Z}$  and an endomorphism  $\omega \in \text{End}_{F_A}(F_{A \cup T})$  satisfying (i).

We start with  $R_0^* = Q^{-1}z_1^{-1}c_1z_1 \dots z_m^{-1}c_mz_m$ ,  $V_0 = \text{Var}(Q)$  and  $\theta_0 = \eta$  given by the hypothesis of the proposition. For  $\phi$  and  $\omega$ , we take the identity maps.

A principal distinction from the proof of Proposition 4.6 is that we cannot use degenerate transformations now. We can have variables  $x$  with  $x^\theta = 1$  which we call *degenerate*. Instead of  $|R[\theta]|$  as one of the inductive parameters, we use  $|\bar{R}[\theta]|$  where  $\bar{R}$  is obtained by removal of all degenerate variables from  $R$  and performing subsequent cancellations. Variables  $x \in \text{Var}(R) \setminus \text{Var}(\bar{R})$  which are not degenerate are called *cancelled*.

Our inductive parameter now is the pair  $(|R^\theta|, |\bar{R}[\theta]|)$  with lexicographic ordering. The whole transformation process consists of steps 1–5 described below.

We start with describing several elementary transformations of triples  $(R^*, V, \theta)$  of the described form. There will be two types of them.

A *substitution* is given by a new set of variables  $V_1 \supseteq V$ , an  $F_A$ -monomorphism  $\psi : F_{A \cup V \cup Z} \rightarrow F_{A \cup V_1 \cup Z}$  and a homomorphism  $\theta_1 : F_{A \cup V_1 \cup Z} \rightarrow F_{A \cup T}$  such that  $\theta = \psi\theta_1$ . The new quadratic word  $R_1^*$  is defined as the cyclically reduced form of  $(R^*)^\psi$ . In most cases when substitutions are defined, we change only  $R$  and do not change the coefficient part  $z_{\sigma(1)}^{-1}c_{\sigma(1)}^{\varepsilon_1}z_{\sigma(1)} \dots z_{\sigma(m)}^{-1}c_{\sigma(m)}^{\varepsilon_m}z_{\sigma(m)}$ .

A *generalization* is given by an endomorphism  $\tau \in \text{End}_{F_A}(F_{A \cup T})$  and a new parametric solution  $\theta_1$  of  $R^* = 1$  such that  $\theta = \theta_1\tau$ . In this case we get a new triple  $(R_1^*, V_1, \theta_1)$  where  $R_1^* = R^*$  and  $V_1 = V$ .

Note that in both cases existence of  $\phi$  and  $\omega$  satisfying (i) for a triple  $(R^*, V, \theta)$  automatically implies one for the new triple  $(R_1^*, V_1, \theta_1)$ . So we will not care about condition (i).

*Transferring cancelled subwords.* Application condition: a word of the form  $x^\varepsilon D$  occurs in  $R$  where  $x \in \text{Var}(\bar{R})$  and  $D$  disappears in  $\bar{R}$ . We apply the substitution  $\psi = (x^\varepsilon \rightarrow x^\varepsilon D^{-1})$  to  $R$  and do not change  $V$  and  $\theta$ . The transformation transfers  $D$  to another location in  $R$ . (Observe that  $x^\varepsilon D x^{-\varepsilon}$  cannot occur in  $R$  since  $x$  would be cancelled otherwise.)

*Cancellation reduction.* Application condition: a non-trivial cancellation in  $\bar{R}[\theta]$  occurs between the values of two neighboring variables  $x^\varepsilon$  and  $y^\delta$  ( $\varepsilon, \delta = \pm 1$ ) and  $x^\varepsilon y^\delta$  occurs also in  $R$ . If  $x \neq y$  then for some  $u, v$  and  $w \neq 1$ ,

$$(x^\varepsilon)^\theta = uw \quad \text{and} \quad (y^\delta)^\theta = w^{-1}v$$

We introduce a new variable  $z \notin V$ , take  $V_1 = V \cup \{z\}$  and define  $\psi$  and  $\theta_1$  by

$$\psi = (x^\varepsilon \mapsto x^\varepsilon z, \quad y^\delta \mapsto z^{-1}y^\delta)$$

and

$$(x^\varepsilon)^{\theta_1} = u, \quad (y^\delta)^{\theta_1} = v, \quad z^{\theta_1} = w \quad \text{and} \quad h^{\theta_1} = h^\theta \text{ for } h \neq x, y, z.$$

The case  $x = y$  is treated in a similar way (see the proof of Proposition 4.6).

Observe that both operations do not change  $R^\theta$ , transferring cancelled subwords does not change also  $\bar{R}[\theta]$  and cancellation reduction decreases the length of  $\bar{R}[\theta]$  by  $2|w|$  where  $w$  is the cancellable part.

*Step 1.* Assume that  $\bar{R}[\theta]$  has a cancellation between the values of two neighboring variables  $x^\varepsilon$  and  $y^\delta$ . Then a word of the form  $x^\varepsilon D y^\delta$  occurs in  $R$  where  $D$  disappears in  $\bar{R}$ . We transfer  $D$  to another location so that  $x^\varepsilon$  and  $y^\delta$  become neighbors in  $R$  and then apply

cancellation reduction decreasing  $|\bar{R}[\theta]|$ . We repeat the procedure until  $\bar{R}[\theta]$  becomes freely reduced.

To make  $\bar{R}[\theta]$  cyclically reduced, we use one more transformation.

*Conjugation.* Take a new variable  $y \notin V$ , take  $V_1 = V \cup \{y\}$  and define  $\psi : F_{A \cup V \cup Z} \rightarrow F_{A \cup V_1 \cup Z}$  by

$$\psi = (x \mapsto y^{-1}xy \text{ for } x \in \text{Var}(R), \quad z_i \rightarrow z_i y \text{ for } i = 1, \dots, m)$$

To define  $\theta_1$ , we take any element  $u \in F_{AUT}$  for the value  $y^{\theta_1}$  of  $y$  and set according to equality  $\theta = \psi\theta_1$ :

$$x^{\theta_1} = ux^{\theta}u^{-1} \text{ for } x \in \text{Var}(R) \quad \text{and} \quad z_i^{\theta_1} = z_i^{\theta}u, \quad i = 1, \dots, m.$$

For this transformation, we have  $R_1^{\theta_1} = u^{-1}R^{\theta}u$ .

*Step 2.* If  $R^{\theta}$  is not cyclically reduced then using conjugation we replace it with its cyclically reduced form. Then jump back to Step 1.

We can assume now that  $R$  and  $\theta$  satisfy condition (ii). In the rest of the proof, we show how to achieve (iii). We observe for the future that using conjugation we can replace  $\bar{R}[\theta]$  with any its cyclic shift not increasing the inductive parameter  $(|R^{\theta}|, |\bar{R}[\theta]|)$ . We introduce yet another transformation.

*Splitting a variable.* Let  $x \in \text{Var}(\bar{R})$ . By the definition of  $\bar{R}$  we have  $x^{\theta} \neq 1$ . Take a new variable  $y \notin V$ , take  $V_1 = V \cup \{y\}$  and apply to  $R$  the substitution  $\psi = (x \mapsto xy)$ . To define  $\theta_1$ , we take any values  $x^{\theta_1}$  and  $y^{\theta_1}$  such that the product  $x^{\theta_1}y^{\theta_1}$  is reduced and equals  $x^{\theta}$ . The values of all other variables are unchanged.

Starting from now we fix any diagram  $\Delta$  with boundary label  $\bar{R}[\theta]$  over the set  $\{c_1, \dots, c_m\}$  folded from the rose diagram associated with factorization (3). Transformations below will include also change of  $\Delta$ .

We introduce a transformation which changes the coefficient part.

*Rearranging coefficients.* Let  $\Delta_0$  be a rose diagram obtained from  $\Delta$  by unfolding, and let

$$\bar{R}[\theta] = v_{\sigma(1)}^{-1}c_{\sigma(1)}v_{\sigma(1)} \cdots v_{\sigma(m)}^{-1}c_{\sigma(m)}v_{\sigma(m)}$$

be the associated factorization of  $\bar{R}[\theta]$  into a product of conjugates of  $c_i$ 's. (Note that we can always assume that a cyclic shift of  $c_i$  coincides with  $c_i$  in this factorization, by performing extra unfolding operations on the rose diagram.)

By Lemma 5.3, there is an automorphism  $\psi \in \text{Aut}(F_{A \cup V \cup Z})$  changing only variables in  $Z$  such that

$$z_1^{-1}c_1z_1 \cdots z_m^{-1}c_mz_m \xrightarrow{\psi} z_{\sigma(1)}^{-1}c_{\sigma(1)}z_{\sigma(1)} \cdots z_{\sigma(m)}^{-1}c_{\sigma(m)}z_{\sigma(m)}$$

and for  $\theta_1 = \psi^{-1}\theta$  we have  $z_i^{\theta_1} = v_i$ . We apply the substitution  $\psi$  to get a new triple  $(R^*, V, \theta_1)$  where the solution  $\theta$  and the coefficient part in  $R^*$  are only changed.

*Step 3.* Let  $e$  be a tree edge of  $\Delta$  and let  $h = \lambda(e) \neq 1$ . The edge  $e$  divides  $\Delta$  into the union  $\Delta = \Delta_1 \cup e \cup \Delta_2$  of  $e$  and two subdiagrams  $\Delta_1$  and  $\Delta_2$ . Passing to a cyclic shift of  $\bar{R}[\theta]$  if needed (which can be performed using conjugation) we assume that

$$\bar{R}[\theta] = uhvh^{-1}$$

where occurrences of  $h$  and  $h^{-1}$  are the labels of  $e$  and  $e^{-1}$ , respectively. Splitting variables if necessary we may assume that  $h$  and  $h^{-1}$  are values of single variables, that is,

$$\bar{R} = Ux_1Vx_2^{\delta}, \quad U^{\theta} = u, \quad x_1^{\theta} = h, \quad V^{\theta} = v \quad \text{and} \quad (x_2^{\delta})^{\theta} = h^{-1}.$$



Note that  $U$  and  $V$  are labels of boundary loops of subdiagrams  $\Delta_1$  and  $\Delta_2$ . The set of 2-cells of  $\Delta$  is partitioned into the set of 2-cells in  $\Delta_1$  and in  $\Delta_2$ . Let

$$U^\theta = w_{i_1}^{-1} c_{i_1} w_{i_1} \dots w_{i_k}^{-1} c_{i_k} w_{i_k}$$

and

$$V^\theta = w_{i_{k+1}}^{-1} c_{i_{k+1}} w_{i_{k+1}} \dots w_{i_m}^{-1} c_{i_m} w_{i_m}$$

where

$$\{1, 2, \dots, m\} = \{i_1, \dots, i_k\} \uplus \{i_{k+1}, \dots, i_m\}$$

be factorizations obtained from unfoldings of  $\Delta_1$  and  $\Delta_2$ . Then

$$\bar{R}[\theta] = (w_{i_1}^{-1} c_{i_1} w_{i_1}) \dots (w_{i_k}^{-1} c_{i_k} w_{i_k}) (h w_{i_{k+1}}^{-1} c_{i_{k+1}} w_{i_{k+1}} h^{-1}) \dots (h w_{i_m}^{-1} c_{i_m} w_{i_m} h^{-1}).$$

is a factorization associated to an unfolding of  $\Delta$ . We apply first rearrangement of coefficients so that equation  $R^* = 1$  gets the form

$$R = EF \quad \text{where} \quad E = z_{i_1}^{-1} c_{i_1} z_{i_1} \dots z_{i_k}^{-1} c_{i_k} z_{i_k}, \quad F = z_{i_{k+1}}^{-1} c_{i_{k+1}} z_{i_{k+1}} \dots z_{i_m}^{-1} c_{i_m} z_{i_m}$$

and

$$E^\theta = U^\theta = u, \quad F^\theta = h V^\theta h^{-1} = h v h^{-1}.$$

Consider two cases.

*Case 1:*  $x_1 = x_2$ . Since  $x_1^\theta = (x_2^{-\delta})^\theta \neq 1$  we have  $\delta = -1$ .

We apply a generalization transformation which replaces  $h$  with a new parameter  $t$ . The new parametric solution  $\theta_1$  is defined by

$$\theta_1 : \begin{cases} x_1 \mapsto t \\ z_j \mapsto z_j^\theta h t^{-1} & \text{for } j = i_{k+1}, \dots, i_m \\ y \mapsto y & \text{for } y \in \text{Var}(R) \setminus \{x_1\} \text{ and } y = z_{i_1}, \dots, z_{i_k} \end{cases}$$

In  $\Delta$ , we change the label of  $e$  to  $t$ . Note that we do not change the inductive parameter  $(|R^\theta|, |\bar{R}[\theta]|)$  since  $h$  is the label of a single edge  $e$  and hence  $|h| = 1$ .

After performing the transformation we start a new iteration of Step 3 checking for another tree edge of  $\Delta$ . In general, after splittings of variables new tree edges may appear in  $\Delta$ . However, the total length of labels of tree edges is not changed. Therefore, after finitely many iteration steps we either process all tree edges of  $\Delta$  or come to Case 2 where the inductive parameter decreases.

*Case 2:*  $x_1 \neq x_2$ . Then  $x_1$  occurs either in  $U$  or in  $V$ . Without loss of generality we assume that  $x_1$  occurs in  $U$  (if  $x_1$  occurs in  $V$  then we can pass to a cyclic shift of  $\bar{R}[\theta]$  by conjugation and then come to a symmetric situation). Let  $U = U_1 x_1^{-1} U_2$ .

First we apply the substitution

$$\psi_1 = (x_1 \mapsto x_1 E^{-1} U_1).$$

We get

$$\begin{aligned} R^{-1} E F &= x_2^{-\delta} V^{-1} x_1^{-1} U_2^{-1} x_1 U_1^{-1} E F \\ &\xrightarrow{\psi_1} x_2^{-\delta} V^{-1} U_1^{-1} E x_1^{-1} U_2^{-1} x_1 F \end{aligned}$$

For the new value  $x_1^{\theta_1}$  of  $x_1$  we have

$$x_1^{\theta_1} = (x_1 U_1^{-1} E)^\theta = U_2^\theta$$

Next we apply another substitution

$$\psi_2 = (z_{i_j} \mapsto z_{i_j} x_1^{-1} U_2 x_1, \quad j = 1, \dots, k)$$

to get

$$x_2^{-\delta} V^{-1} U_1^{-1} E x_1^{-1} U_2^{-1} x_1 F \xrightarrow{\psi_2} x_2^{-\delta} V^{-1} U_1^{-1} x_1^{-1} U_2^{-1} x_1 E F.$$

After application of  $\psi_1$  and  $\psi_2$  we get a new triple  $(R_1^*, V, \theta_1)$  with  $R_1 = x_1^{-1} U_2 x_1 U_1 V x_2^\delta$  and

$$R_1^{\theta_1} = (U_2 U_1 V x_2^\delta)^\theta$$

which implies

$$|R_1^{\theta_1}| < |R^\theta|.$$

We jump next to Step 1. Step 3 is finished.

At this point, we produce a triple  $(R^*, V, \theta)$  satisfying (ii) and a diagram  $\Delta$  with boundary label  $\bar{R}[\theta]$  over the set  $\{c_1, \dots, c_m\}$  folded from the rose diagram associated with (3). The diagram  $\Delta$  has the property that the label of any its tree edge  $e$  is a single parameter letter  $t$ . Moreover, there are exactly two occurrences of  $t$  in  $\bar{R}[\theta]$  and both are the values of one variable  $x \in \text{Var}(\bar{R})$ .

*Step 4.* Let  $e$  be a tubular edge of  $\Delta$ . Let  $D$  be the trivial 2-cell with boundary loop  $epe^{-1}q$  where  $p$  and  $q$  are loops with empty labels and  $p$  bounds a subdiagram  $\Delta_1$  of  $\Delta$ . Without loss of generality we assume that there are no tubular edges in  $\Delta_1$ . We choose any non self-intersecting path  $s$  joining the base vertex  $\nu_0$  with the start of  $e$  and unfold  $\Delta$  along the path  $sepe^{-1}s^{-1}$  as shown in Fig. 10 (note that  $D$  is contracted into an edge after this procedure). Then we perform folding of the resulting diagram reducing cancellation in the newly appeared copies  $s_1 e_1 p_1 e_1^{-1} s_1^{-1}$  and  $s_2 e_2 p_2 e_2^{-1} s_2^{-1}$  of  $sepe^{-1}s^{-1}$  (Fig. 10). The resulting diagram  $\Delta'$  is the union of two subdiagrams  $\Theta$  and  $\Delta_2$  with  $\Theta \cap \Delta_2 = \{\nu_0\}$ . The subdiagram  $\Theta$  is obtained from the union of  $\Delta_1$  and  $s_1 e_1$  by adding trivial 2-cells so that the complement  $\Theta - \Delta$  consists of annuli formed by the trivial 2-cells and tubular edges that become all edges of the path  $s_1 e_1$ . The subdiagram  $\Delta_2$  is obtained from the complement  $\Delta - \Delta_1$  by contracting the path  $q$  into a vertex (to do the contraction, we introduce trivial edges where needed to remove self-intersections in  $q$ ).

Observe that the boundary label of  $\Theta$  is empty and the boundary label of  $\Delta'$  is equal to the boundary label of  $\Delta$ . Since both  $\Delta$  and  $\Delta'$  can be unfolded form a common rose diagram  $\Delta_0$  we can perform rearranging coefficients so that factorization (3) is replaced by the factorization associated to  $\Delta_0$ . We replace also  $\Delta$  with  $\Delta'$ .

Similar to Step 3 we further replace occurrences of the label of  $s_1 e_1$  in the values  $x^\beta$  of variables  $x$  by a new parameter letter (only values of variables  $z_i$  are changed for which the corresponding coefficient  $c_i$  is the boundary label of a 2-cell in  $\Delta_1$ ).

We perform the described procedure for all tubular edges of  $\Delta$ . After this,  $\Delta$  becomes the union of subdiagrams  $\hat{\Delta}$  and  $\Theta_1, \dots, \Theta_r$  with a common vertex  $\nu_0$  and having no other intersections. Each subdiagram  $\Theta_i$  has empty label and a single tubular edge labelled by a parameter letter. It is not hard to see that after the whole procedure no tree edges appear in  $\Delta$ .

*Step 5.* The final procedure is elimination of tree vertices of  $\Delta$ , that is, vertices that do not belong to the boundary of any 2-cell of  $\Delta$ .

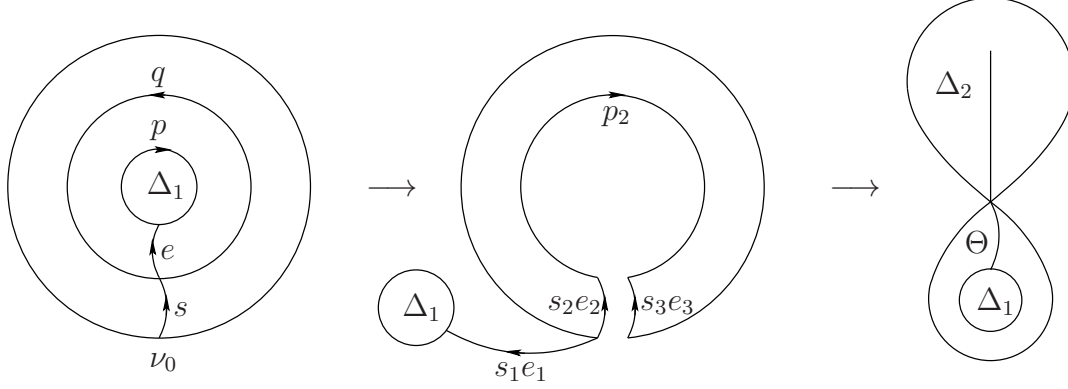


FIGURE 10.

Let  $\nu$  be a tree vertex of  $\Delta$ . Since  $\bar{R}[\theta]$  is cyclically reduced, the valence of  $\nu$  is at least 2. We assume without loss of generality that  $\nu$  is distinct from the base vertex  $\nu_0$  of  $\Delta$  (otherwise using conjugation we can move  $\nu_0$  to any non-tree boundary vertex of  $\Delta$ ).

Let  $e_1, \dots, e_d$  be all directed edges starting at  $\nu$ , and let  $x_1, \dots, x_d$  be the corresponding variables, so  $x_i^\theta = t_i$  is a parameter letter which is the label of  $e_i$ . Since  $\nu \neq \nu_0$ , all occurrences of variables  $x_i$  in  $\bar{R}$  are of the form  $\dots x_i^{-1} x_{i+1} \dots$  ( $i \bmod d$ ). Hence all occurrences of  $x_i$  in  $R$  have the form  $\dots x_i^{-1} D_i x_{i+1} \dots$  ( $i \bmod d$ ) where all  $D_i$  disappear in  $\bar{R}$ . We first apply to  $R$  the substitution

$$(x_2 \mapsto D_1^{-1} x_1 x_2, x_3 \mapsto D_2^{-1} D_1^{-1} x_1 x_3, \dots, x_k \mapsto D_k^{-1} \dots D_1^{-1} x_1 x_k)$$

This eliminates  $x_2$  from  $\bar{R}$  (it either disappears in  $R$  or becomes cancelled in  $\bar{R}$ ). For the new solution  $\theta_1$  we get  $x_i^{\theta_1} = t_1 t_i$  for  $i = 2, \dots, d$ . All occurrences of parameter letters  $t_i$  in the values of other variables (which can be only  $z_j$ 's) are of the form  $\dots t_i^{-1} t_{i+1} \dots$  ( $i \bmod d$ ). Hence we can perform a generalization replacing each  $t_1 t_i$  with a single parameter  $t_i$ .

After performing this operation, the length of  $\bar{R}[\theta]$  is decreased. We get also a new diagram  $\Delta_1$  which is obtained from  $\Delta$  by contracting the edge  $e_1$  into a vertex.

We repeat the procedure until we get rid off all tree vertices of  $\Delta$ .

The description of the transformation process is finished. Let  $\Delta$  be a diagram obtained after all steps 1–5. To prove the proposition, we have only to estimate the total number of tree and tubular edges of  $\Delta$ .

Let  $e_1, \dots, e_k$  be all tree edges of  $\Delta$  and let  $r$  be the number of subdiagrams  $\Theta_i$  with empty label produced at Step 4. Each of the subdiagrams  $\Theta_i$  has at least one non-trivial 2-cell since otherwise it should be contracted to a vertex. The same is true for all connected components of the complement  $\Delta - \cup_i e_i$ . This implies that the number of these connected components is at most  $m - r$ . On the other hand, this number is precisely  $k + 1$  since each  $e_i$  joins two components of  $\Delta - \cup_i e_i$  and  $\Delta$  is simply connected. Hence  $k + r < m$ .

The proof of Proposition 6.2 is completed in the case when  $Q$  is orientable. In the case of non-orientable  $Q$  the only difference in the argument is that we admit inverses of coefficient factors. (In Case 2 at Step 3, it may happen that  $x_1$  occurs in  $R$  twice with the same exponent  $+1$  or  $-1$ . Then the corresponding transformation inverses  $E$  which occurs in the coefficient part.)  $\square$

*Proof of Theorem 1.2.* We start with the case when  $Q$  is a standard orientable quadratic word.

We assume that  $Q$  has at least one coefficient. If  $Q$  is coefficient-free then there exists a parametric solution  $\eta_0 \in \text{Hom}(F_{\text{Var}(Q)}, F_T)$  of the equation  $Q = 1$  such that the value of each variable is either a parameter letter or trivial and any ordinary solution  $\alpha$  of  $Q = 1$  in *any* coefficient group  $F_A$  may be represented as  $\phi\alpha\omega$  where  $\phi \in \text{Stab}(Q)$  and  $\omega \in \text{Hom}(F_T, F_A)$  (see Theorem 4 in Section 5 of a survey [12]). This implies that  $\eta_0$  is a generalization of any other parametric solution  $\eta$  of  $Q = 1$  (since  $\eta$  is an ordinary solution in  $F_{A \cup T}$ ). In this case, the statement of Theorem 1.2 holds with bound  $|x^{\eta_0}| \leq 1$ .

Let  $\eta$  be a parametric solution of a standard quadratic equation  $Q = 1$  where

$$Q = [x_1, y_1][x_2, y_2] \dots [x_g, y_g] c_1 z_2^{-1} c_2 z_2 \dots z_m^{-1} c_m z_m$$

Similarly to the proof of Theorem 1.1 we may assume that all  $c_i$  are cyclically reduced. We adjust  $Q$  and  $\eta$  by introducing an extra variable  $z_1$  with  $z_1^\eta = 1$  and replacing  $c_1$  in  $Q$  with  $z_1^{-1} c_1 z_1$ . By Proposition 6.2 with  $Q := ([x_1, y_1] \dots [x_g, y_g])^{-1}$  we find an equation of the form

$$R = z_{\sigma(1)}^{-1} c_{\sigma(1)} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)} z_{\sigma(m)},$$

a finite set of variables  $V \supseteq \text{Var}(R)$  and a parametric solution  $\theta : F_{A \cup V \cup Z} \rightarrow F_{A \cup T}$  which satisfy conditions (i)–(iii) of that proposition.

Let a primitive  $F_A$ -monomorphism  $\phi : F_{A \cup \text{Var}(Q_0) \cup Z} \rightarrow F_{A \cup V \cup Z}$  and  $\omega \in \text{End}_{F_A}(F_{A \cup T})$  be as in (i), that is, up to conjugation we have

$$(4) \quad Q \xrightarrow{\phi} R^{-1} z_{\sigma(1)}^{-1} c_{\sigma(1)} z_{\sigma(1)} \dots z_{\sigma(m)} c_{\sigma(m)} z_{\sigma(m)}$$

and the following diagram is commutative:

$$(5) \quad \begin{array}{ccc} F_{A \cup \text{Var}(Q_0) \cup Z} & \xrightarrow{\phi} & F_{A \cup V \cup Z} \\ \eta \downarrow & & \downarrow \theta \\ F_{A \cup T} & \xleftarrow{\omega} & F_{A \cup T} \end{array}$$

Our strategy is to further transform the equation  $R = z_{\sigma(1)}^{-1} c_{\sigma(1)} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)} z_{\sigma(m)}$  and its parametric solution  $\theta$  so that we come back to the initial equation  $Q = 1$ . After that, with a slight adjustment,  $\theta$  will give a desired generalization of  $\eta$ . We will find an “ecomonic” transformation which, together with conditions (ii) and (iii) of Proposition 6.2, will provide the required bound on the size of the resulting generalization of  $\eta$ .

As in the proof of Proposition 6.2, each moment we have the following data:

- A quadratic word  $R^*$  of the form

$$R^* = R^{-1} z_{\sigma(1)}^{-1} c_{\sigma(1)} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)} z_{\sigma(m)}$$

where  $R$  is a coefficient-free quadratic word equivalent to  $[x_1, y_1] \dots [x_g, y_g]$ ;

- A parametric solution  $\theta : F_{A \cup V \cup Z} \rightarrow F_{A \cup T}$  of the equation  $R^* = 1$ .
- A primitive  $F_A$ -monomorphism  $\phi : F_{A \cup \text{Var}(Q) \cup Z} \rightarrow F_{A \cup V \cup Z}$  and an endomorphism  $\omega \in \text{End}_{F_A}(F_{A \cup T})$  with (4) and (5).

Note that we do not include the set  $V$  of formal variables here since we do not need to introduce new variables and  $V$  will not change.

In a similar way, we will use two types of elementary transformations: substitutions and generalizations. During the transformation process, condition (i) of Proposition 6.2 will be automatically held by construction.

We start now with the pair  $(R^*, \theta)$  obtained after application of Proposition 6.2. Observe that conditions (ii) and (iii) imply the following bound on the total length of values  $x^\theta$  of variables in  $x \in \text{Var}(\bar{R})$ :

$$\sum_{x \in \text{Var}(\bar{R})} |x^\theta| \leq \frac{1}{2}c(Q) + m.$$

*Step 1: Transforming the coefficient part.* Let  $\Delta$  be the diagram satisfying condition (iii) of Proposition 6.2. We unfold  $\Delta$  using Proposition 5.4 and in a similar way as in the proof of Theorem 1.1 we change the coefficient part of equation  $R^* = 1$  and the parametric solution  $\theta$  using Proposition 5.3 so that the equation gets the form

$$R = z_1^{-1}c_1z_1 \dots z_m^{-1}c_mz_m$$

and we have

$$|z_i^\theta| \leq 3c(Q) + 4m \quad \text{for all } i = 1, 2, \dots, m.$$

*Step 2: Transforming  $\bar{R}$  to the standard form.* We apply the procedure described in the proof of Proposition 3.9 to make  $\bar{R}$  a product of commutators. Since our transformation should apply to  $R$  we mimic application of related Nielsen automorphisms to  $\bar{R}$  as application of automorphisms to  $R$  in the following way.

Assume that  $x^\varepsilon y^\delta$  occurs in  $\bar{R}$  and we want to apply to  $\bar{R}$  a related Nielsen automorphism  $\rho = (x^\varepsilon \rightarrow x^\varepsilon y^{-\delta})$ . There is a subword  $x^\varepsilon W y^\delta$  of  $R$  where  $W$  is deleted in  $\bar{R}$ . In particular,  $W^\theta = 1$ . Then application of an automorphism  $(x^\varepsilon \rightarrow x^\varepsilon y^{-\delta} W^{-1})$  to  $R$  produces a new word  $R_1$  such that  $\bar{R}_1 = \bar{R}^\rho$ .

At this step, we change the values of parametric solution  $\theta$  on variables of  $x \in \text{Var}(\bar{R})$  only (and will not change them until the final Step 4). According to Lemma 3.12 we get

$$|x^\theta| \leq 2c(Q) + 4m \quad \text{for all } x \in \text{Var}(\bar{R}).$$

*Step 3: Transforming the deleted part to the standard form.* Recall that variables in  $\text{Var}(R) \setminus \text{Var}(\bar{R})$  are divided into two types: *degenerate* variables  $x$  with  $x^\theta = 1$  and *cancelled* ones which cancel in  $R$  after removal of degenerate variables.

Let

$$R = W_0 x_1^{\varepsilon_1} W_1 x_2^{\varepsilon_2} \dots W_{k-1} x_k^{\varepsilon_k} W_k$$

where  $\bar{R} = x_1^{\varepsilon_1} x_2^{\varepsilon_2} \dots x_k^{\varepsilon_k}$  and  $W_i$  are deleted in  $\bar{R}$ . Using substitutions  $(x_i^{\varepsilon_i} \rightarrow x_i^{\varepsilon_i} W_i^{-1})$  and  $(x_{i+1}^{\varepsilon_{i+1}} \rightarrow W_i^{-1} x_{i+1}^{\varepsilon_{i+1}})$  we can move  $W_i$ 's not changing  $\bar{R}$ . Since  $\bar{R}$  is a product of commutators, it is easy to see that we can collect all  $W_i$ 's together at any location between  $x_i^{\varepsilon_i}$  and  $x_{i+1}^{\varepsilon_{i+1}}$ . We reduce  $R$  to the form

$$R = W \bar{R}.$$

Let  $x, y \in \text{Var}(W)$  and  $\sigma = (x^\varepsilon \mapsto x^\varepsilon y^\delta) \in \text{Aut}(F_{A \cup V \cup Z})$  be a Nielsen automorphism related to  $R$  (i.e.  $(x^\varepsilon y^{-\delta})^{\pm 1}$  occurs in  $W$ ). Application of  $\sigma$  does not change  $\bar{R}$  whenever  $x$  is cancelled or both  $x$  and  $y$  are degenerate. Observing that two cancelled variables cannot “cross” in  $W$  and using Nielsen automorphisms of the above form we transform  $W$  to a product of commutators

$$[y_1, y_2] \dots [y_{2r-1}, y_{2r}]$$

where at least one variable  $y_j$  or  $y_{j+1}$  is degenerate in each commutator  $[y_i, y_{i+1}]$ .

Finally, we apply a generalization to  $\theta$  by assigning a single parameter letter to each cancelled variable  $x$ . After the transformation, all of  $R$  is written as a product of commutators and we get

$$|x^\theta| \leq 1 \quad \text{for all } x \in \text{Var}(R) \setminus \text{Var}(\bar{R})$$

*Step 4: Eliminating  $z_1$ .* We do this in the same way as in the proof of Theorem 1.1. The resulting bound increases by  $6c(Q) + 8m$  for  $|x^\theta|$  when  $x \in \text{Var}(R)$  and by  $3c(Q) + 4m$  for  $|z_i^\theta|$ .

We have “almost” produced the desired generalization  $\theta$  of the initial parametric solution  $\eta$ , with the only difference that  $\theta$  is formally defined on a larger set of variables  $V$ . This issue is solved by the following observation.

**Lemma 6.3.** *Let  $Q \in F_{A \cup X}$  be a standard quadratic word with at least one coefficient. If  $\phi : F_{A \cup \text{Var}(Q)} \rightarrow F_{A \cup V}$  is a primitive  $F_A$ -monomorphism and  $Q^\phi$  is conjugate to  $Q$  then  $\text{Im } \phi = F_{A \cup \text{Var}(Q)}$ .*

*Proof.* It follows from Definition 6.1 that a primitive  $F_A$ -monomorphism  $\phi : F_{A \cup \text{Var}(Q)} \rightarrow F_{A \cup V}$  can be extended to an  $F_A$ -automorphism of  $F_{A \cup V}$ . Assume that  $\phi \in \text{Aut}_{F_A}(F_{A \cup V})$  and  $Q^\phi$  is conjugate to  $Q$ . We apply a slightly modified version of Proposition I.4.24 in [21] to one cyclic word  $u_1 = Q$  and the tuple of non-cyclic words  $u_i$ ,  $i \geq 2$ , consisting of all letters of  $A$ . Observe that there is no elementary Whitehead automorphism  $\rho \in \text{Aut}(F_{A \cup V})$  which does not increase the length of all  $u_i$  and strictly decreases it for at least one  $i$ . Then  $\phi$  can be represented as a product  $\rho_1 \rho_2 \dots \rho_r$  of elementary Whitehead automorphisms  $\rho_i$  which do not change the length of each  $u_i$ . We can eliminate permutations and exponent sign changes of generators, so each  $\rho_i$  can be assumed to be an elementary Whitehead  $F_A$ -automorphism of  $F_{A \cup V}$ . An easy inductive argument shows that then for each  $i$ , the image  $Q^{\rho_1 \dots \rho_i}$  (viewed as a cyclic word) is a quadratic word equivalent to  $Q$  and  $(F_{A \cup \text{Var}(Q)})^{\rho_i} = F_{A \cup \text{Var}(Q)}$ . This implies that  $(F_{A \cup \text{Var}(Q)})^\phi = F_{A \cup \text{Var}(Q)}$ .  $\square$

The lemma shows that as long as we have a commutative diagram (5) where  $\phi \in \text{Stab}(Q)$  then then we can restrict it to  $F_{A \cup \text{Var}(Q)}$ :

$$\begin{array}{ccc} F_{A \cup \text{Var}(Q)} & \xrightarrow{\tilde{\phi}} & F_{A \cup \text{Var}(Q)} \\ \eta \downarrow & & \downarrow \tilde{\theta} \\ F_{A \cup T} & \xleftarrow{\omega} & F_{A \cup T} \end{array}$$

This means precisely that  $\tilde{\theta}$  is a generalization of  $\eta$ . The proof of Theorem 1.2 is finished in the case of a standard orientable quadratic equation  $Q = 1$ .

The case when  $Q$  is semi-standard non-orientable is treated in a similar way with the following changes:

- The word  $R^*$  under transformation has the form

$$R^* = R^{-1} z_{\sigma(1)}^{-1} c_{\sigma(1)}^{\varepsilon_1} z_{\sigma(1)} \dots z_{\sigma(m)}^{-1} c_{\sigma(m)}^{\varepsilon_m} z_{\sigma(m)}.$$

- After Step 1, the equation gets the form

$$R = z_1^{-1} c_1^{\varepsilon_1} z_1 \dots z_m^{-1} c_m^{\varepsilon_m} z_m$$

for some  $\varepsilon_1, \dots, \varepsilon_m = \pm 1$ .

- In Step 2, we reduce  $R$  to a semi-standard form.



- In Step 3, we reduce  $W$  to a semi-standard form

$$y_1^2 \cdots y_k^2 [y_{k+1}, y_{k+2}] \cdots [y_{r-1}, y_r]$$

where each  $y_i$  in  $y_i^2$  is degenerate and either  $y_i$  or  $y_{i+1}$  is degenerate in each commutator  $[y_i, y_{i+1}]$ . After the reduction, we move  $W$  to the right of the last square factor in  $\bar{R}$  so that the whole word  $R$  would be written in a semi-standard form.

- Before Step 4, we apply Lemma 3.14 to get  $R^* = Q$ .

We prove the theorem in the general case. Let  $Q = 1$  be any quadratic equation in  $F_A$ .

Let  $\phi \in \text{Aut}_{F_A}(F_{A \cup \text{Var}(Q)})$  be an automorphism given by Propositions 3.3 or 3.5 such that  $Q^\phi$  is conjugate to a standard or semi-standard quadratic word  $R$  equivalent to  $Q$ . Then  $\phi$  defines a one-two-one correspondence  $\eta \mapsto \phi\eta$  between parametric solutions  $\eta$  of the equation  $(R = 1, \text{Var}(Q))$  with formal set of variables  $\text{Var}(Q)$  and parametric solutions  $\phi\eta$  of the equation  $Q = 1$ .

Recall that a parametric solution  $\eta_1 : F_{A \cup \text{Var}(E)} \rightarrow F_{AUT}$  of  $E = 1$  is a generalization of another parametric solution  $\eta_2$  of the same equation  $E = 1$  if there are an automorphism  $\psi \in \text{Stab}(E)$  and an endomorphism  $\omega \in \text{End}_{F_A}(F_{AUT})$  such that  $\eta_2 = \psi\eta_1\omega$ . We extend this definition to parametric solutions of equations  $(E = 1, V)$  with formal set of variables by taking instead of  $\text{Stab}(E)$  the group

$$\text{Stab}(E, V) = \{\psi \in \text{Aut}_{F_A}(F_{A \cup V}) \mid E^\psi \text{ is conjugate to } E\}.$$

Clearly, the correspondence  $\eta \mapsto \phi\eta$  preserves the relation “ $\eta_1$  is a generalization of  $\eta_2$ ” in the new extended version.

For a parametric solution  $\theta$  of  $R = 1$ , let  $\hat{\theta}$  denote the parametric solution of  $(R = 1, \text{Var}(Q))$  defined by extending  $\theta$  in the following natural way: for each variable  $x \in \text{Var}(Q) \setminus \text{Var}(R)$  we choose a parameter letter  $t$  (which does not occur in parametric words  $y^\theta$  for  $y \in \text{Var}(R)$ ) and set  $x^{\hat{\theta}} = t$ . Clearly, any parametric solution of  $(R = 1, \text{Var}(Q))$  has a generalization of the form  $\hat{\theta}$  for some  $\theta$ . It is also obvious that the correspondence  $\theta \mapsto \hat{\theta}$  preserves the relation “ $\theta_1$  is a generalization of  $\theta_2$ ”.

Now let  $\eta$  be an arbitrary parametric solution of  $Q = 1$ . To find the required generalization  $\eta_1$  of  $\eta$ , we pass on to the parametric solution  $\psi^{-1}\xi$  of  $(R = 1, \text{Var}(Q))$  and take its generalization of the form  $\hat{\theta}$  for some  $\theta$ . By the statement of the theorem for the equation  $R = 1$ , there is a generalization  $\theta_1$  of  $\theta$  such that for any  $x \in \text{Var}(R)$

$$|x^{\theta_1}| \leq \begin{cases} 8(c(R) + 2n(R)) & \text{if } Q \text{ is orientable,} \\ 36n(R)(c(R) + 2n(R)) & \text{if } Q \text{ is non-orientable.} \end{cases}$$

We take  $\eta_1 = \psi\hat{\theta}_1$ . By the bounds on  $\psi$  in Propositions 3.3 or 3.5 (and using inequalities  $n(R) \leq n(Q)$  and  $c(R) \leq c(Q)$ ), for any  $x \in \text{Var}(Q)$  we have

$$\begin{aligned} |x^{\eta_1}| &\leq 4 \sum_{x \in \text{Var}(Q)} |x^{\hat{\theta}_1}| + 2c(Q) \\ &\leq \begin{cases} 34n(R)(c(R) + 2n(R)) & \text{if } Q \text{ is orientable,} \\ 146n(R)^2(c(R) + 2n(R)) & \text{if } Q \text{ is non-orientable.} \end{cases} \end{aligned}$$

as required. □

## REFERENCES

- [1] D. Bormotov, R. Gilman and A. Myasnikov, *Solving one-variable equations in free groups*, J. Group Theory 12 (2009), no. 2, 317–330.
- [2] G. Baumslag, A. Myasnikov, V. Remeslennikov, *Algebraic geometry over groups I. Algebraic sets and ideal theory*, J. Algebra, 219, 1999, 1, 16–79.
- [3] I. Bumagin, O. Kharlampovich and A. Miasnikov, *Isomorphism problem for finitely generated fully residually free groups*, J. Pure Appl. Algebra, 208 (2007), 961–977.
- [4] L. Comerford, *Quadratic equations over small cancellation groups*, J. Algebra, 69, 1981, 1, 175–185.
- [5] L. Comerford and C. Edmunds, *Quadratic equations over free groups and free products*, J. Algebra, 68, 1981, 2, 276–297.
- [6] L. Comerford and C. Edmunds, *Solutions of equations in free groups*, Group Theory (Singapore 1987), 347–356, de Gruyter, Berlin, 1989.
- [7] M. Culler, *Using surfaces to solve equations in free groups*, Topology, 20, 1981, 2, 133–145.
- [8] F. Dahmani, D. Groves, *The isomorphism problem for toral relatively hyperbolic groups*, Publ. IHES 107 (2008), 211–290.
- [9] F. Dahmani, V. Guirardel, *The isomorphism problem for all hyperbolic groups*, ArXiv:1002.250v2
- [10] V. Diekert, J. Michael, *Quadratic word equations*, Jewels are forever, 1999, Springer, Berlin, 314–326.
- [11] R. Grigorchuk and P. Kurchanov, *On quadratic equations in free groups*, Proceedings of the International Conference on Algebra, Part 1 (Novosibirsk, 1989), 1992, Contemp. Math. 131, 159–171.
- [12] R. Grigorchuk and P. Kurchanov, *Some questions of group theory related to geometry*, A. N. Parshin, I. R. Shafarevich (Eds.), Algebra VII. Combinatorial group theory. Applications to geometry. Springer 1993. 167–232.
- [13] O. Kharlampovich, I. G. Lysënok, A. G. Myasnikov and N. W. M. Touikan, *The Solvability Problem for Quadratic Equations over Free Groups is NP-Complete*, Theory Comput. Syst. 47 (2010), no. 1, 250–258.
- [14] O. Kharlampovich, A. Myasnikov, *Implicit function theorem over free groups*, J. Algebra, 290, 2005, 1, 1–203.
- [15] O. Kharlampovich, A. Myasnikov, *Irreducible affine varieties over a free group. I. Irreducibility of quadratic equations and Nullstellensatz*, J. Algebra, 200, 1998, 2, 472–516.
- [16] O. Kharlampovich, A. Myasnikov, *Irreducible affine varieties over a free group. II. Systems in triangular quasi-quadratic form and description of residually free groups*, J. Algebra, 200, 1998, 2, 517–570.
- [17] O. Kharlampovich, A. Myasnikov, "Equations and algorithmic problems in groups", publicacoes matematicas, IMPA, Brasil, 2008.
- [18] O. Kharlampovich, A. Myasnikov, *Equations and fully residually free groups*, Combinatorial and Geometric Group Theory. Dortmund and Carleton conferences (2007), New Trends in Mathematics, Birkhauser, 2010, 203–243.
- [19] O. Kharlampovich, A. Myasnikov, *Elementary theory of free nonabelian groups*. Journal of Algebra, 2006, Volume 302, Issue 2, p. 451–552.
- [20] R.C. Lyndon, *The equation  $a^2b^2 = c^2$  in free groups*, Michigan Math. J., 6 (1959) 155–164.
- [21] R.C. Lyndon, P. Schupp, *Combinatorial group theory*, Springer, 1977.
- [22] A. I. Mal'cev, *On the equation  $zxyx^{-1}y^{-1}z^{-1} = aba^{-1}b^{-1}$  in a free group*, (Russian) Algebra i Logika Sem., 1, 1962 no. 5, 45–50.
- [23] G. Makanin, *Equations in a free group*, Izv. Akad. Nauk SSSR Ser. Mat., 46, 1982, 6, 1199–1273.
- [24] A. Ol'shanskii, *Diagrams of homomorphisms of surface groups*, Sibirsk. Mat. Zh., 30, 1989, 6, 150–171.
- [25] A. Razborov, *On systems of equations in a free group*, Candidate dissertation, Steklov Math. Institute, 1987.
- [26] Seifert and Threlfall, *A textbook of topology*, Academic Press, 1980.
- [27] Z. Sela, *The isomorphism problem for hyperbolic groups. I*. Ann. of Math. 141 (1995), no. 2, 217–283.
- [28] Z. Sela, *Diophantine geometry over groups I–VI*, Publ. Math. IHES 93 (2001), 31–105, Israel J. Math. 134 (2003), 173–254, Israel J. Math. 143 (2004), 1–130, Israel J. Math. 147 (2005), 1–73, Israel J. Math. 150 (2005), 1–197, Geom. Funct. Anal. 16 (2006), no. 3, 537–706, Geom. Funct. Anal. 16 (2006), no. 3, 707–730.

- [29] M. Wicks, *Commutators in free products*, J. London Math. Soc., 37, 1962, 433–444.
- [30] M. Wicks, *A general solution of binary homogeneous equations over free groups*, Pacific J. Math., 41, 1972, 543–561.

STEKLOV MATHEMATICAL INSTITUTE, MOSCOW, RUSSIA

*E-mail address:* igor.lysenok@gmail.com

DEPARTMENT OF MATHEMATICS, STEVENS INSTITUTE OF TECHNOLOGY, HOBOKEN, USA

*E-mail address:* amiasnikov@gmail.com